**Universidade do Minho**
Escola de Engenharia

Patrícia Dinis Mota da Costa

# Modelos Multidimensionais de Resposta ao Item

Março de 2011

Patrícia Dinis Mota da Costa **Modelos Multidimensionais de Resposta ao Item**
UMinho | 2011

**Universidade do Minho**
Escola de Engenharia

Patrícia Dinis Mota da Costa

# Modelos Multidimensionais de Resposta ao Item

Programa Doutoral em Engenharia Industrial e de Sistemas

Trabalho efectuado sob a orientação da
**Professora Doutora Maria Eugénia Ferrão**
e do
**Professor Doutor Pedro Nuno Oliveira**

Março de 2011

# Agradecimentos

Quero expressar os meus sinceros agradecimentos:

- À Professora Doutora Maria Eugénia Ferrão, pela orientação, dedicação, sugestões e pelos conselhos preciosos que foram sempre acompanhados de exigência científica, por ter revelado uma atenção e disponibilidade inesgotáveis na concretização deste trabalho e pela amizade patenteada;

- Ao Professor Doutor Pedro Oliveira, pela orientação, disponibilidade, dedicação, pelo apoio e pelos estímulos durante a realização desta tese;

- Ao Departamento de Matemática da Universidade da Beira Interior, através do projecto de investigação intitulado "Eficácia Escolar no Ensino da Matemática", pelo financiamento da bolsa de investigação e por me ter proporcionado as condições necessárias para a elaboração deste trabalho.

- Aos meus Pais, pela formação humana que me deram e pelas oportunidades que me proporcionaram a todos os níveis. Pelo seu carinho e amor;

- Ao meu irmão pelo encorajamento que sempre me transmitiu e apoio incondicional, sempre presentes durante a realização deste trabalho;

- Ao Jorge, por todo o carinho, pela infinita paciência, ajuda, compreensão e pelas palavras certas;

- Aos meus sobrinhos, Rafael e David, pela força que me deram e pela compreensão dos momentos em que não pudemos estar juntos;

- A todos os que de alguma forma contribuíram para a consecução deste trabalho.

# Modelos Multidimensionais de Resposta ao Item

# Resumo

Neste trabalho estudamos as propriedades dos modelos de resposta ao item em termos teóricos e de aplicabilidade a dados reais. Em particular: exploramos modelos de resposta ao item unidimensionais dicotómicos e politómicos; aplicamos modelos unidimensionais para grupos múltiplos; utilizamos procedimentos estatísticos de equalização e *linking* para comparar classificações obtidas pela aplicação de diferentes instrumentos e generalizamos os modelos unidimensionais logísticos de 1, 2 e 3 parâmetros a modelos multidimensionais.

Propomos estimar os parâmetros dos itens e dos factores latentes do modelo multidimensional de resposta ao item compensatório logístico de 2 parâmetros, conjugando a estimação bayesiana com o uso de métodos de simulação *Markov Chain Monte Carlo* (MCMC). Para isso, recorremos ao algoritmo de *Metropolis-Hastings* com amostragem *Gibbs*. A estimação de todos os parâmeros do modelo é feita simultaneamente. Para testar o procedimento, usamos dados simulados considerando 2 e 3 factores latentes. Utilizamos o critério de informação de Akaike (AIC) para seleccionar o número de dimensões que melhor se adequa aos dados. Os resultados mostram que se obtêm boas estimativas pela aplicação do procedimento proposto em termos de correlação, de erro absoluto médio e de erro quadrático médio. Com o propósito de verificarmos os resultados obtidos, aplicamos a abordagem proposta a dados reais, recolhidos no âmbito do projecto de investigação Eficácia Escolar no Ensino da Matemática (3EM). A metodologia adoptada é inovadora e os resultados obtidos confirmam a sua apropriação para a estimação dos parâmetros do modelo.

# Multidimensional Item Response Models

# Abstract

In this work we propose to study psychometric properties of item response models in theoretical terms and in the applicability to real data. In particular, we explore unidimensional dichotomous and polytomous item response models; we apply unidimensional item response model to multiple groups; we use the statistical procedures equating and linking to compare results obtained by the application of different instruments and we generalize unidimensional 1, 2 and 3 parameter logistic item response models to multidimensional models.

We propose to estimate item parameters and latent factors of the 2 parameter logistic multidimensional compensatory item response model, using bayesian estimation procedure and simulation methods, Markov Chain Monte Carlo (MCMC). In order to do this, we use the Metropolis-Hastings algorithm with steps of Gibbs. The estimation of all parameters of the model is done simultaneously. To test the procedure, we use simulated data considering 2 and 3 latent factors. To select the number of dimensions that best fit the data, we utilize the Akaike's information criteria (AIC). The results show that good estimates are obtained by the proposed procedure in terms of correlation, mean absolute error and root mean square error. With the propose to confirm the results obtained, we apply the proposed procedure to real data, collected as part of a research project entitled School Effectiveness in Mathematics Teaching (3EM). The methodology adopted is an innovation and the results confirm the appropriation to the estimates of the model parameters.

# Conteúdo

# Lista de Tabelas

# Lista de Figuras

# Introdução

O investimento dos países em educação constitui uma das mais importantes vertentes de desenvolvimento. Este investimento permitiu que, ao longo do tempo, fossem criados métodos estatísticos mais eficientes, possibilitando a obtenção de resultados, que, por sua vez, impulsionaram a implementação dos chamados Sistemas de Avaliação em Larga Escala. Nesse sentido, a utilização de Modelos de Resposta ao Item (MRI) foi de muita importância, pelo conjunto de características particulares que apresentam, permitindo obter resultados que anteriormente eram praticamente impossíveis, como por exemplo, a comparação de resultados de grupos de examinandos submetidos a instrumentos diferentes. Esta classe de modelos é usada para analisar dados provenientes de respostas a itens que constituem instrumentos de avaliação de habilidades, questionários, entre outros. Os MRI relacionam formas de representar a probabilidade de um examinando responder correctamente a um item tendo em conta os seus factores latentes/habilidades, na área de conhecimento avaliada, e as características/propriedades do item. Actualmente, esta classe de modelos é utilizada em diversos países mostrando-se como um instrumento poderoso nos processos quantitativos de avaliação educacional, das quais se destacam as seguintes aplicações: elaboração de instrumentos, criação de escalas de medida, construção de bancos de itens, utilização de testes sob medida e investigação de propriedades dos itens.

A classe de modelos de resposta ao item foi explorada na dissertação de Mestrado efectuada e intitulada "Modelos de Resposta ao Item" (Costa [24]). Nesta disser-

tação, foram introduzidos os principais conceitos, modelos e resultados que se obtêm a partir da aplicação de modelos de resposta ao item unidimensionais. Nesse sentido, foram apresentadas as especificações formais dos modelos de resposta ao item, assumindo os pressupostos de unidimensionalidade do factor latente e de uma população subjacente, de itens na forma dicotómica (certo/errado). Os três modelos especificados foram: modelo logístico de 1 parâmetro, modelo logístico de 2 parâmetros e modelo logístico de 3 parâmetros. Foram discutidas algumas propriedades desses modelos e expostos os principais procedimentos de estimação, utilizando o método da Máxima Verosimilhança Marginal (MVM) de Bock e Lieberman [18] para: (1) a estimação dos parâmetros dos itens considerando o factor latente conhecido; (2) a estimação do factor latente considerando os parâmetros dos itens conhecidos; (3) a estimação conjunta dos parâmetros dos itens e do factor latente. Para este método foram descritos o algoritmo de Newton - Raphson e o método *scoring* de Fisher. Adicionalmente, foram apresentadas abordagens recorrendo a métodos iterativos, ao método da quadratura, à abordagem Bock e Aitkin [16] e ao algoritmo *Expectation-Maximization* (EM). Foram realizadas duas aplicações desta classe de modelos: a criação de uma escala de qualidade da infra-estrutura das escolas (Costa e Ferrão [25]) e a criação de uma escala de desempenho em Matemática. Nesta última aplicação foi efectuada a selecção, a alteração e a remoção de itens, com vista à construção de um banco de itens para aferir o desempenho em Matemática nos vários anos de escolaridade do Ensino Básico (Ferrão *et al.* [44]). Para tal, foram estabelecidos critérios de entrada e saída de itens com base na análise da função de informação do item e da sua curva característica.

No decorrer da dissertação foram ainda identificadas "oportunidades" de investigação para futuro desenvolvimento. Este trabalho surge no seguimento das linhas de investigação futuras da dissertação e tem como objectivos:

1) explorar modelos de resposta ao item politómicos unidimensionais;

2) incorporar modelos que englobam a comparação de grupos diferentes de

examinandos - conhecidos como modelos para grupos múltiplos;

3) aplicar os procedimentos estatísticos de equalização e *linking*;

4) generalizar os modelos unidimensionais logísticos de 1, 2 e 3 parâmetros a modelos multidimensionais;

5) explorar procedimentos de simulação, *Markov Chain Monte Carlo* (MCMC), para a optimização dos procedimentos de estimação. Em particular, propor um procedimento de estimação bayesiano, através do uso de MCMC, para estimar os parâmetros dos modelos multidimensionais de resposta ao item.

Os modelos de resposta ao item unidimensionais politómicos são importantes, na medida em que incorporam várias categorias de respostas (além das dicotómicas), o que promove a versatilidade das suas aplicações. Os modelos para grupos múltiplos permitem estudar examinandos que provêm de diferentes grupos/populações. A aplicação de procedimentos de equalização e *linking* engloba desafios práticos em situações em que se pretende comparar resultados.

Estas linhas de investigação têm sido desenvolvidas nos últimos anos, das quais resultaram publicações em periódicos nacionais e internacionais. Entre 2007 e 2010 participei na equipa de investigação no Laboratório de Estatística Aplicada e Computacional do Departamento de Matemática da Universidade da Beira Interior, num projecto de investigação intitulado "Eficácia Escolar no Ensino da Matemática - 3EM[1]". Neste projecto de investigação tive como principais actividades: o estudo das propriedades estatísticas de instrumentos de medida dos resultados escolares, utilizando de modo complementar as abordagens da Teoria Clássica dos Testes (TCT) e modelos de resposta ao item; a análise da qualidade de itens de instrumentos de aferição de aprendizagens; análise e selecção de itens; construção de escalas de desempenho; criação de banco de itens para aferir aprendizagens a Matemática

[1]Informações mais detalhadas sobre o projecto 3EM podem ser encontradas no Anexo 3 - secção 3.3.

(Ferrão *et al.* [44]); a comparação de resultados escolares utilizando procedimentos de equalização (Costa, Oliveira e Ferrão [28]) e a análise da dimensionalidade de testes pelos métodos de informação restrita e os métodos de informação plena (Costa *et al.* [26]). Participei na execução dos relatórios intitulados "Provas de Aferição de Matemática, Português do 4º e 6º anos de escolaridade"[98] e "Testes 3EMat e Provas de Aferição (Ligação entre escalas de desempenho, considerando a disciplina de Matemática do 6º ano de escolaridade no ano lectivo 2006/7)"[99], no âmbito do projecto Melhoria da Qualidade dos Instrumentos e Escalas de Aferição dos Resultados Escolares[2], desenvolvido em parceria entre o Gabinete de Avaliação Educacional (GAVE) do Ministério da Educação e a Universidade da Beira Interior (UBI). No primeiro relatório foi explorada a utilização do MRI politómico de Crédito Parcial Generalizado e no segundo relatório foi usado o procedimento *linking* para a comparação dos instrumentos.

No âmbito de um projecto da Universidade do Minho que visa a utilização, de forma complementar, da TCT e dos MRI para garantir a qualidade na aferição das aprendizagens na unidade curricular de Estatística do curso de Mestrado Integrado em Engenharia e Gestão Industrial[3], foram estudadas as propriedades psicométricas de testes de escolha múltipla. Com o propósito de melhorar os itens, apresentámos os piores itens de um teste de Estatística descritiva e analisámo-los em termos de objectivos específicos de aprendizagem (Costa, Oliveira e Ferrão [30]).

No que se refere a procedimentos de estimação de parâmetros de MRI, Costa, Fletcher e Ferrão [27] usaram o algoritmo EM no procedimento de estimação de MVM para estimar os parâmetros do MRI unidimensional logístico de 2 parâmetros, com uma abordagem semelhante à proposta por Bock e Aitkin em 1981 [16]. Esta abordagem recorre ao teorema de Bayes para determinar a distribuição condicional do factor latente dadas as respostas aos itens do teste. O algoritmo proposto foi desenvolvido em linguagem R. Ferrão, Costa e Gama [46] apresentaram uma abor-

[2]No Anexo 3 - secção 3.2 podem ser encontradas informações mais detalhadas sobre o projecto.

[3]Mais detalhes podem ser consultados no Anexo 3 - secção 3.1.

dagem metodológica baseada no uso de onduletas, para estimar a função densidade do factor latente do modelo unidimensional logístico de 2 parâmetros, comparativamente com a assunção do pressuposto de uma distribuição normal. Esta abordagem é uma extensão do procedimento de MVM e foi implementada em linguagem R. A análise dos resultados (erro absoluto médio - EAM, erro quadrático médio - EQM e testes de qualidade de ajuste) sugere que se verificam melhores resultados em termos de menores erros e de maior probabilidade de não rejeitar a hipótese nula quando as onduletas são consideradas.

Adicionalmente, têm sido apresentadas diversas comunicações, relacionadas com esta temática, em congressos nacionais (Sociedade Portuguesa de Estatística - SPE e Encontro de Economia Econometria e Métodos Estatísticos em Educação - CEMAPRE) e internacionais (*International Conference on Teaching Statistics* - ICOTS8, *European Society for Engineering Education* - SEFI, Simpósio Nacional de Probabilidade e Estatística - SINAPE, Congresso Brasileiro de Teoria da Resposta ao Item - CONBRATRI, Sociedade Brasileira de Pesquisa Operacional - SOBRAPO e Associação Brasileira de Avaliação Educacional - ABAVE).

Nos modelos multidimensionais o pressuposto da existência de apenas um factor latente é abandonado e assim, estes modelos surgem para superar algumas das limitações dos modelos unidimensionais, nomeadamente: ignorarem as correlações entre os factores latentes e fornecerem medidas imprecisas quando os testes têm um reduzido número de itens. Na estimação dos parâmetros dos modelos multidimensionais, as limitações apontadas para a utilização de procedimentos de máxima verosimilhança, relacionam-se com o facto de não estarem definidos para alguns padrões de resposta, como itens com acerto total ou erro total e respostas omissas; com a possibilidade das estimativas dos parâmetros dos itens não pertencerem ao intervalo esperado; com a ocorrência frequente de problemas numéricos em testes longos e com a dificuldade na aplicação do algoritmo EM, quando as configurações do teste são mais complexas. Já o uso de procedimentos de estimação bayesianos em modelos de

resposta ao item exige cálculos mais complexos do que os procedimentos de máxima verosimilhança. Nesse sentido, começaram a ser utilizados métodos de simulação com vista a obter as estimativas dos parâmetros em MRI. A utilização de métodos MCMC possibilita o uso de simulações para obter as estimativas dos parâmetros de modelos de resposta ao item, pela aplicação procedimentos de estimação bayesianos, o que permite diminuir as dificuldades na estimação destes parâmetros tanto a nível teórico como a nível computacional. Em particular, neste trabalho usamos procedimentos de simulação MCMC para estimar os parâmetros do modelo multidimensional compensatório logístico de 2 parâmetros. O algoritmo que propomos permite obter, simultaneamente, as estimativas dos parâmetros dos itens e dos factores latentes dos examinandos. Para isso, utilizamos o algoritmo de *Metropolis-Hastings* com amostragem *Gibbs*. O algoritmo foi implementado em Matlab. Inicialmente, usamos dados simulados para testar o procedimento de estimação proposto, considerando que os dados aferem 2 e 3 factores latentes. Adicionalmente, aplicamos o procedimento de estimação bayesiano proposto a dados reais e comparamos os resultados obtidos com os do software comercial Testfact (Wilson, Wood e Gibbons [115]), que recorre ao procedimento de estimação de máxima verosimilhança para estimar os parâmetros do modelo, possibilitando assim a comparação entre os vários procedimentos de estimação.

Esta tese focada no estudo e aplicação de modelos de resposta ao item unidimensionais e multidimensionais, está organizada em 4 capítulos. O capítulo 1, Modelos Unidimensionais de Resposta ao Item, compreende 9 secções. Nesse sentido, começamos por fazer uma breve introdução a esta classe de modelos. Nas secções 1.2 e 1.3, abordamos os postulados e os pressupostos desta classe de modelos. Seguidamente, efectuamos a especificação formal dos modelos dicotómicos actualmente existentes. Na secção 1.5, debruçamo-nos sobre os modelos de resposta ao item politómicos. Na secção seguinte exploramos os modelos para grupos múltiplos. Nas secções 1.7 e 1.8, apresentamos os conceitos de função de informação do item e

de função de informação do teste, respectivamente. Na última secção deste capítulo, descrevemos os procedimentos de equalização e *linking* em termos da sua revisão da literatura, dos modos de desenvolvimento dos testes, dos planos de recolha de dados e dos principais métodos utilizados para efectuar a comparação/ligação entre os instrumentos.

No capítulo 2, Modelos Multidimensionais de Resposta ao Item, começamos por efectuar uma introdução a esta classe de modelos, onde apresentamos a revisão da literatura dos modelos multidimensionais em termos da generalização de modelos unidimensionais. Na secção seguinte, especificamos formalmente os principais modelos multidimensionais. Na secção 2.3, debruçamo-nos sobre os procedimentos de estimação. Esta secção está organizada da seguinte forma: apresentamos, inicialmente, os procedimentos de estimação mais utilizados e as suas limitações, seguidamente descrevemos os procedimentos de simulação de MCMC em MRI e para finalizar propomos o procedimento de estimação inovador para a estimação dos parâmetros dos modelos multidimensionais que conjuga procedimentos bayesianos com MCMC. Nas considerações gerais do capítulo, expomos os dois métodos utilizados para a análise de dimensionalidade de instrumentos: método de análise factorial de informação restrita e método de análise factorial de informação plena e discutimos o que foi apresentado no capítulo.

No capítulo 3, Aplicações - Modelos de resposta ao item unidimensionais, debruçamo-nos sobre as aplicações desta classe de modelos. Este capítulo divide-se em 4 aplicações: modelos para dados dicotómicos, modelos para dados politómicos, modelos para grupos múltiplos e procedimentos de equalização e *linking*.

Efectuamos as aplicações dos modelos de resposta ao item multidimensionais no capítulo 4. Este compreende as seguintes aplicações: análise de dimensionalidade de um teste de Matemática; utilização de modelos de resposta ao item, usando o procedimento de estimação proposto a dados simulados, e a aplicação desta classe de modelos a dados reais.

Nas conclusões e trabalhos futuros, discutimos os resultados apurados, apresentamos as conclusões deste trabalho de investigação, indicamos as limitações encontradas ao longo do mesmo e referimo-nos também a alguns trabalhos futuros, relacionados com os assuntos em estudo, possíveis de uma exploração mais detalhada.

# Capítulo 1

# Modelos Unidimensionais de Resposta ao Item

## 1.1 Introdução

Na avaliação educacional e psicológica, os testes são instrumentos que permitem aferir conhecimentos nas diversas áreas. Existem duas abordagens principais para o estudo da qualidade da avaliação baseada em testes: a Teoria Clássica dos Testes (Lord e Novick [78]) e Modelos de Resposta ao Item (Hambleton, Swaminathan e Rogers [59]).

Na TCT as respostas aos itens são consideradas certas ou erradas e a soma das respostas correctas é o resultado do teste. Os resultados encontrados dependem de um conjunto particular de itens que compõem o instrumento de medida, ou seja, as análises e interpretações estão sempre associadas como um todo. Em geral, nesta abordagem, a caracterização das propriedades de um teste é feita com recurso às estatísticas: índice de discriminação, índice de dificuldade e correlação ponto bisserial.

Supõe-se que a variável $V$ representa o conhecimento/habilidade do examinando. A realização da variável $V$ é, usualmente, obtida através da aplicação de um teste

donde se obtém o resultado do examinando no mesmo. Considera-se que a habilidade do examinando é o que está a ser aferido no teste (Boring [22]). Se os testes fossem instrumentos de medição com precisão absoluta, o valor obtido, $V^0$, por aplicação de qualquer deles, seria igual ao valor verdadeiro, $V$. Na situação hipotética em que a habilidade do examinando é testada $T$ vezes, o modelo seguinte representa a relação entre o valor verdadeiro da variável e o respectivo valor observado,

$$V_t^0 = V + \epsilon_t \tag{1.1.1}$$

com $(t = 1, .., T)$ e onde $\epsilon_t$ representa o erro da medição.

Assume-se que o erro é não sistemático e homocedástico, isto é, $E[\epsilon_t] = 0$ e $Var[\epsilon_t] = \sigma^2$. Adicionalmente, assume-se que o erro é não correlacionado com o valor verdadeiro, pelo que $E[V_t^0] = E[V_t]$ e $Var[V_t^0] = Var[V] + Var[\epsilon_t]$. O pressuposto de que o erro segue uma distribuição normal é necessário ao processo inferencial.

As características dos itens, tais como a capacidade de discriminação e a dificuldade são quantificadas através dos respectivos índices de discriminação e de dificuldade. Adicionalmente, a correlação ponto-bisserial quantifica a associação entre o item e $V_0$.

O índice de discriminação (por exemplo, Guilford e Fruchter [55]) mede a capacidade do item diferenciar os examinandos com alto desempenho (27% dos examinandos com resultados mais altos) daqueles que têm baixo desempenho (27% dos examinandos com resultados mais baixos). Este índice obtém-se pela diferença entre a proporção de acerto no item dos examinandos que tiveram classificação superior ao percentil 73 (grupo de alto desempenho) e a proporção de acerto no item dos examinandos com classificação inferior ao percentil 27 (grupo de baixo desempenho). Os valores obtidos para este parâmetro variam entre -1 e 1. Os itens que apresentam índices de discriminação superiores a 0,4 são considerados muito discriminativos, os que têm valores que pertencem ao intervalo [0,3; 0,4] são designados discriminativos e com valores inferiores a 0,3 são considerados pouco discriminativos.

O índice de dificuldade é dado pela proporção de acerto no item. Portanto,

valores altos para este índice indicam que os itens são fáceis. Os índices de dificuldade podem ser classificados em 5 categorias da seguinte forma: muito difícil ([0;0,25[), difícil ([0,25;0,45[), médio ([0,45;0,55[), fácil ([0,55;0,75[) e muito fácil ([0,75;1]).

O coeficiente de correlação ponto-bisserial mede a correlação do resultado de um item em particular do teste com o resultado do teste como um todo, sendo, portanto, uma medida da capacidade de discriminação do item relativamente ao resultado total do teste. Este coeficiente obtém-se pelo cálculo do coeficiente de Bravais-Pearson (D'Hainaut [36]) e é dado por:

$$r_{pbi} = \frac{M_1 - M_0}{s}\sqrt{p(1-p)} \tag{1.1.2}$$

onde

$M_1$ representa a média do resultado dos examinandos que acertaram o item;

$M_0$ representa a média do resultado dos examinandos que erraram o item;

$p$ representa a proporção de examinandos que acertaram o item.

$s$ representa o desvio padrão do resultado no teste de todos os examinandos.

Os itens que apresentem valores para esta estatística inferiores a 0,2 devem ser revistos ou até retirados do teste. É de notar que alguns autores usam a correlação bisserial (D'Hainaut [36]).

A fiabilidade do procedimento de medição é um indicador do erro de medição e o coeficiente de fiabilidade é usado para a sua quantificação (Hand [60]), sendo definido pela expressão de cálculo seguinte:

$$R = \frac{Var[V]}{Var[V_t^0]} = 1 - \frac{Var[\epsilon_t]}{Var[V] + Var[\epsilon_t]} \tag{1.1.3}$$

Deste modo, $0 \leq R \leq 1$, quando $R = 1$ o valor observado é o valor verdadeiro.

Considerando a natureza dos itens, isto é, o tipo de respostas aos itens serem dicotómicas ou se distribuirem por uma escala ordinal, o estimador para a fiabilidade é dado pela correlação de Kuder-Richardson KR20 [72](Dunn [42]) ou pelo coeficiente *alpha* de Cronbach [31], respectivamente.

Apesar das potencialidades da TCT, esta teoria apresenta algumas limitações teóricas, nomeadamente:

1) os parâmetros dos itens (índice de dificuldade e índice de discriminação) dependem directamente da amostra de examinandos para estabelecê-los;

2) os testes/instrumentos são dependentes dos itens que os compõem, o que significa que testes diferentes que afiram a mesma habilidade produzem resultados diferentes, para examinandos com as mesmas características;

3) a TCT é orientada para o teste global e não para o item individual. Toda a informação do item deriva de considerações do teste geral, não se podendo assim, determinar como o examinando se comportaria perante cada item individual;

4) quando se pretende comparar resultados da aferição da habilidade ao longo de anos sucessivos, a TCT impõe a utilização de testes paralelos[1] ou do mesmo teste.

Para superar algumas das limitações que a TCT continha, e pela necessidade de procurar metodologias alternativas para colmatar as dificuldades dos modelos e das técnicas clássicas de medida, surgiram os MRI. Daqui em diante, quando as habilidades não são observadas directamente designam-se por factores latentes. Os MRI constituem uma classe de modelos estatísticos que representam a relação entre a probabilidade de um examinando responder correctamente a um item e o seu(s) factor(es) latente(s) na área do conhecimento avaliada. Essa relação é sempre expressa de tal forma que quanto maior o(s) factor(es) latente(s) maior a probabilidade de acerto no item.

Segundo Baker [8], os MRI surgiram a partir dos trabalhos de Lord [76] e Rasch [95] e caracterizam-se pela independência do teste aplicado, pela independência da

[1] Dois testes dizem-se paralelos quando aferem a mesma habilidade mas são compostos por itens diferentes.

amostra de examinandos a que é aplicado e pela comparabilidade dos resultados obtidos para amostras de examinandos diferentes, mesmo quando os testes aplicados são parcialmente distintos. Assim, nos MRI os parâmetros dos itens e do factor latente são considerados invariantes. A utilização desta classe de modelos permite realizar a análise de cada item que constitui o instrumento de avaliação ou medida, considerando as suas características na produção das estimativas do factor latente, facilitando, também, a interpretação da escala produzida. Pressupõe-se que a unidade de análise é o item e não o instrumento como um todo, como acontece na TCT.

A classificação dos vários MRI propostos na literatura depende fundamentalmente de três características (Andrade, Tavares e Valle [5]):

1) natureza do item - dicotómicos ou politómicos;

2) número de populações envolvidas - grupo único ou múltiplos grupos;

3) número de factores latentes que está a ser medido - unidimensional ou multidimensional.

Seguidamente, para os MRI unidimensionais, apresentamos os postulados, os pressupostos e os modelos actualmente mais utilizados. Abordamos, depois, a função de informação do item e a função de informação do teste. No final do capítulo, exploramos os procedimentos de equalização e *linking* em termos da sua evolução histórica, de desenvolvimento de testes, dos planos de recolha de dados e dos métodos e procedimentos estatísticos mais utilizados.

## 1.2 Postulados

Os MRI baseiam-se em dois postulados (Hambleton, Swaminathan e Rogers, [59]):

1) O desempenho de um examinando num item do teste é função do factor latente;

2) A relação entre o desempenho no item e o factor latente pode ser descrita por uma função monótona crescente, chamada Curva Característica do Item (CCI). Esta função estabelece que à medida que o nível do factor latente aumenta, a probabilidade de uma resposta correcta ao item também aumenta. A CCI tem inclinação e deslocamento na escala do factor latente definidos pelos parâmetros do item (figura 1.1).

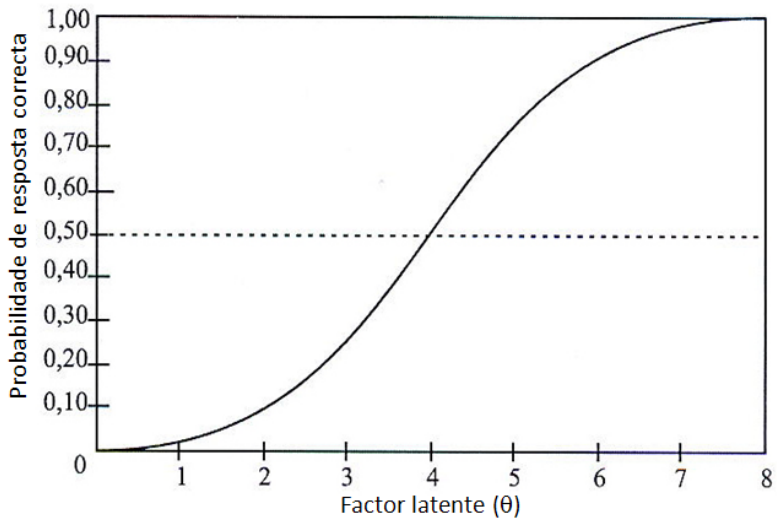



Figura 1.1: Curva característica do item

## 1.3 Pressupostos

As vantagens da utilização dos MRI unidimensionais dependem, fundamentalmente, da verificação dos seus pressupostos. Assim, esta classe de modelos baseia-se em dois pressupostos: unidimensionalidade e independência local.

1) Unidimensionalidade - Os MRI postulam que existe apenas um factor latente responsável pela realização de um conjunto de itens de um instrumento. É de notar que em geral, se considera que, até certo ponto, alguns factores cognitivos, de personalidade e do próprio instrumento podem afectar o desempenho

no instrumento. Contudo, para satisfazer o pressuposto da unidimensionalidade, é necessário assumir que existe um factor latente dominante. Este factor é a habilidade que se supõe estar a ser medida pelo instrumento.

2) Independência Local - A independência local assume que, para um dado factor latente, as respostas aos diferentes itens são independentes. Isto significa que, para examinandos com um dado factor latente, a probabilidade de resposta a um conjunto de itens é igual ao produto das probabilidades das respostas do examinando a cada item individual. Sejam $U_{ij}$ a variável aleatória que representa a resposta dicotómica do examinando $j$ ao item $i$ ($U_{ij} = 1$, se for resposta correcta; $U_{ij} = 0$, se for resposta incorrecta) e $P(U_{ij}|\theta_j)$ a probabilidade de resposta do examinando $j$ dado o seu factor latente $\theta_j$. Considerando os $I$ itens do teste, a probabilidade conjunta dado o $\theta_j$ do examinando é:

$$P(U_{1j}, U_{2j}, ..., U_{Ij}|\theta_j) = \prod_{i=1}^{I} P(U_{ij}|\theta_j) \tag{1.3.1}$$

onde $j = 1, 2, ..., J$.

Segundo Lord [77] e Lord e Novick [78], a unidimensionalidade implica a independência local, já que a única causa da resposta do examinando é o pressuposto do factor latente dominante. Assim, ao verificar-se a unidimensionalidade, a independência local fica subjacente. Os procedimentos estatísticos que se utilizam para verificar o pressuposto da unidimensionalidade vão ser descritos no capítulo 2, secção 2.4.

## 1.4 Modelos dicotómicos

Existem vários MRI que diferem no tipo de função matemática utilizada para definir a CCI e/ou no número de parâmetros especificados no modelo. Todos os modelos unidimensionais consideram um ou mais parâmetros para descrever o item e um

parâmetro para descrever o examinando. Em geral, são utilizadas duas funções matemáticas para caracterizar os parâmetros dos itens: a logística e a ogiva normal. Ambas as funções fornecem informações sobre os parâmetros dos itens através das CCIs. O modelo de ogiva normal apresenta a CCI ligeiramente mais acentuada que a obtida pela aplicação do modelo logístico para os mesmos parâmetros do item. Com vista a tornar os modelos idênticos, Birnbaum [14] propôs multiplicar os expoentes do modelo logístico pela constante $D = 1,702$. Haley [57], nesse caso, mostrou que a função de distribuição normal e a função logística diferem em probabilidade menos de 0,01. A função logística é matematicamente mais conveniente porque é uma função explícita dos parâmetros do item e do factor latente e não envolve integração.

Os modelos unidimensionais mais utilizados, para itens dicotómicos, são os que consideram a função logística. Estes modelos são designados por modelos logísticos de 1, 2 e 3 parâmetros. Os parâmetros que se usam para descrever os itens são:

i) a dificuldade do item - no caso do modelo logístico de 1 parâmetro;

ii) a dificuldade do item e a discriminação do item - no caso do modelo logístico de 2 parâmetros;

iii) a dificuldade do item, a discriminação do item e a probabilidade de resposta correcta dada por examinandos com baixo factor latente (designada por probabilidade de acerto ao acaso) - no caso do modelo logístico de 3 parâmetros.

Nas subsecções seguintes, apresenta-se a descrição de cada um dos modelos referidos.

### 1.4.1 Modelo logístico de 1 parâmetro

O modelo logístico de 1 parâmetro (ML1) é o modelo teórico mais simples. Foi proposto por Rasch [95] (também é denominado por Modelo de Rasch) e a probabilidade de um examinando responder correctamente a um item é dada por:

$$P(U_{ij} = 1|\theta_j) = \frac{e^{D(\theta_j - b_i)}}{1 + e^{D(\theta_j - b_i)}} \tag{1.4.1}$$

onde $i = 1, 2, ..., I$ e $j = 1, 2, ..., J$, com:

$U_{ij}$ é uma variável dicotómica, definida como na secção 1.3;

$\theta_j$ representa o factor latente do $j$-ésimo examinando;

$P(U_{ij} = 1|\theta_j)$ é a probabilidade de um examinando $j$ com factor latente $\theta_j$ responder correctamente ao item $i$;

$b_i$ é o parâmetro de dificuldade do item $i$, medido na mesma escala do factor latente;

$e$ é a base dos logaritmos neperianos (cujo valor é aproximadamente 2,72);

$D = 1,702$;

$I$ é o número de itens do teste;

$J$ é o número de examinandos.

$P(U_{ij} = 1|\theta_j)$ produz a CCI conforme a figura 1.2.

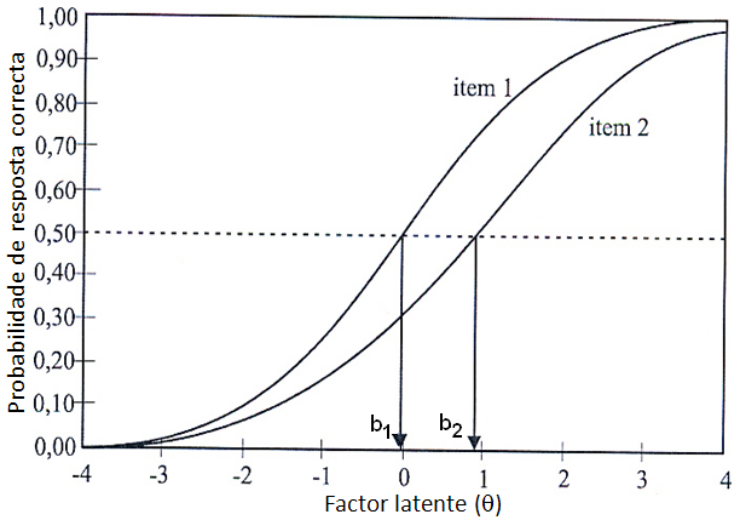



Figura 1.2: O parâmetro de dificuldade de dois itens

O parâmetro $b_i$ é o parâmetro de dificuldade do item e corresponde ao ponto na escala do factor latente em que o examinando alcança 0,5 de probabilidade de responder correctamente ao item $i$. É um parâmetro de localização que indica a

posição na CCI em relação à escala do factor latente. Quanto maior for o $b_i$, maior deve ser o nível do factor latente exigido para que o examinando tenha 50% de possibilidade de acertar no item. Itens difíceis estão localizados à direita ou em valores mais altos na escala do factor latente e itens fáceis estão situados à esquerda ou em valores mais baixos na escala do factor latente. Quando os valores do factor latente de um grupo são transformados para a escala de média igual a 0 e desvio padrão igual a 1, os valores de $b_i$ situam-se, tipicamente, entre $-3$ (itens fáceis) e $+3$ (itens difíceis). Em geral, a classificação da dificuldade dos itens é feita do seguinte modo: os itens que apresentem parâmetro de dificuldade superiores a 0,75 são considerados difíceis, os itens que têm valores que pertencem ao intervalo [-0,75; 0,75] são considerados de dificuldade média e os que têm valores inferiores a -0,75 são considerados fáceis.

### 1.4.2 Modelo logístico de 2 parâmetros

No modelo logístico de dois parâmetros (ML2), a probabilidade de um examinando responder correctamente a um item depende da dificuldade e discriminação. Esta relação é expressa por:

$$P(U_{ij} = 1|\theta_j) = \frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}} \tag{1.4.2}$$

onde

$a_i$ é o parâmetro de discriminação do item $i$;

$P(U_{ij} = 1|\theta_j)$, $b_i$, $e$, $D$, $I$ e $J$ assumem o mesmo significado do ML1.

A diferença deste modelo relativamente com o ML1 está no aparecimento, na sua formulação, do índice de discriminação do item $i$, $a_i$. Este parâmetro indica a inclinação da curva no ponto de inflexão, onde a probabilidade de resposta é 0,5. Itens com maior inclinação são mais úteis para distinguir examinandos com diferentes níveis de factor latente do que itens com menor inclinação. O parâmetro $a_i$ pode variar de 0 a $\infty$, mas tipicamente varia entre 0 e 3. Em geral, considera-se que

valores do parâmetro de discriminação inferiores a 0,4 indicam que os itens são pouco discriminativos; itens que apresentam valores para este parâmetro que pertencem ao intervalo [0,4; 0,7] são considerados discriminativos e itens muito discriminativos são os que têm valores do parâmetro de discriminação superiores a 0,7.

A figura 1.3 mostra a representação gráfica dos parâmetros de dificuldade e discriminação dos itens.

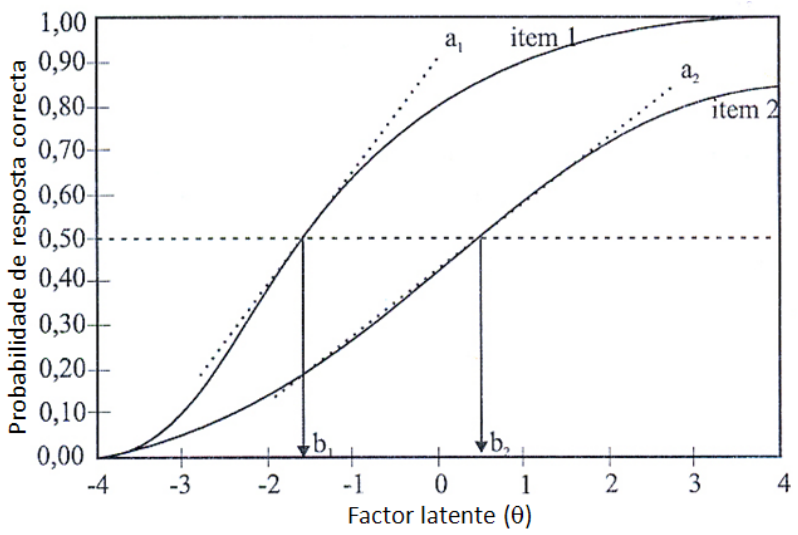



Figura 1.3: Os parâmetros de dificuldade e de discriminação de dois itens.

A análise da figura permite constatar que o item 2 ($b_2$) é mais difícil do que o item 1 ($b_1$) e que o item 2 é menos discriminativo ($a_2$) do que o item 1 ($a_1$), uma vez que a inclinação da curva do item 2 é menor. Adicionalmente, verifica-se que os valores de discriminação diferem pelo facto de as curvas características dos itens não serem paralelas e, consequentemente, apresentarem inclinação diferente.

### 1.4.3 Modelo logístico de 3 parâmetros

O modelo logístico de três parâmetros (ML3) assume que a probabilidade de acerto a um item depende da sua dificuldade, discriminação e probabilidade de acerto ao acaso. A sua formulação matemática apresenta-se seguidamente:

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}} \tag{1.4.3}$$

com

$c_i$ é o parâmetro do item que representa a probabilidade de examinandos com baixo factor latente responderem correctamente ao item $i$ (muitas vezes referido como a probabilidade de acerto ao acaso);

$P(U_{ij} = 1|\theta_j)$, $b_i$, $a_i$, $e$, $D$, $I$ e $J$ assumem o mesmo significado dos modelos definidos anteriormente.

O parâmetro $c_i$ é expresso pela assímptota inferior da CCI e assume sempre valores entre 0 e 1. Se esta assímptota cortar a ordenada acima do ponto 0, há presença de acertos ao acaso. Quanto menor o valor deste parâmetro, melhor é o item. No caso em que a alternativa de resposta correcta é assinalada ao acaso, o valor de $c_i$ é igual a 1/(número de alternativas de resposta ao item).

Em particular, no ML3, o parâmetro $b_i$ representa o ponto na escala do factor latente, para o qual a probabilidade de resposta correcta é $\frac{1+c_i}{2}$. A sua interpretação é análoga à apresentada no ML1. Em termos do parâmetro de discriminação, a maior inclinação da CCI é proporcional a este parâmetro no ponto onde a probabilidade de acerto é de $\frac{1+c_i}{2}$.

A representação gráfica de um exemplo de CCI do ML3, bem como a indicação dos parâmetros dos itens estão apresentadas na figura 1.4.

A figura anterior permite verificar que no item 2, há 0,2 de probabilidade que o item seja acertado ao acaso ($c_2$), sendo esta probabilidade zero para os outros dois itens.

### 1.4.4 Outros modelos

Não faz parte do objectivo deste trabalho apresentar uma discussão de todos os modelos dicotómicos existentes, contudo, nesta subsecção apresenta-se o modelo baseado na função ogiva normal (Lord e Novick [78]), dada a sua importância no

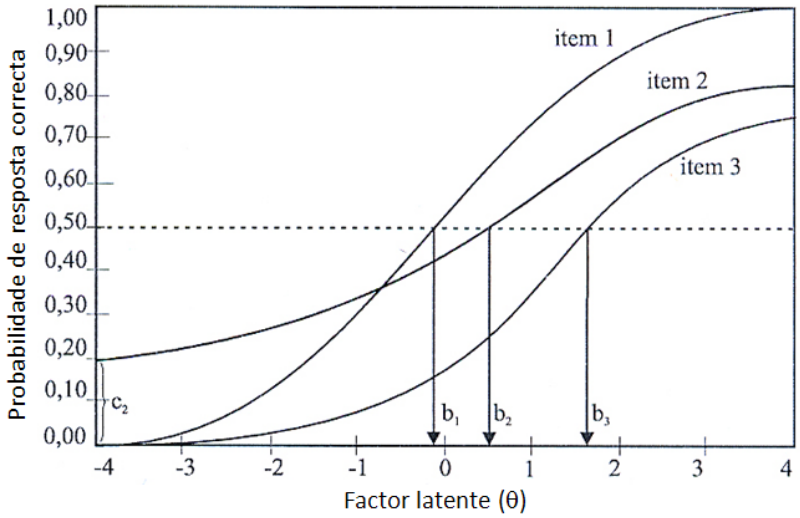



Figura 1.4: CCI de um item obtida pela aplicação do ML3

desenvolvimento dos modelos multidimensionais.

A função ogiva normal equivalente ao ML3 é dada por:

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i) \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \tag{1.4.4}$$

onde $z = a_i(\theta_j - b_i)$ e todos os parâmetros dos itens e do factor latente dos examinandos têm as mesmas definições do que as dos modelos logísticos.

O integral especificado no modelo define a área sob a distribuição normal padrão de $-\infty$ até $z$.

Para além dos modelos dicotómicos apresentados anteriormente, existem muitos outros modelos em que se consideram outras funções matemáticas (Reckase [97]), como: funções lineares (Lazarsfeld [73]), funções polinomiais (McDonald [84], Samejima e Livingston [103], Sympson [110]) e funções *spline* (Abrahamowicz e Ramsay [1]).

Os modelos dicotómicos consideram somente as respostas aos itens classificadas como correcta ou incorrecta. Nesse sentido, surgiram os modelos politómicos com vista a obter mais informação a partir de respostas de examinandos, a itens que podem ser respondidos numa escala ordenada com mais de duas alternativas de

resposta ou onde há respostas parciais ou com créditos parciais. Esta classe de modelos vai ser apresentada na secção seguinte.

## 1.5 Modelos politómicos

Nos modelos politómicos, os itens são variáveis categóricas que incorporam mais do que duas categorias resposta. Este tipo de modelos engloba tanto a análise de itens de resposta aberta como a análise de itens de resposta fechada (ou de escolha múltipla), que são elaborados ou corrigidos de forma a obterem-se categorias intermédias ordenadas entre as categorias certo ou errado. A principal vantagem apontada para o seu uso, é a obtenção de maior quantidade de informação a partir das respostas dos examinandos e pela possibilidade de aferir factor latente que esteja parcialmente desenvolvido.

Uma apresentação detalhada dos modelos politómicos encontra-se em van der Linden e Hambleton [74] e em Andrade, Tavares e Valle [5]. Seguidamente, apresentamos os modelos politómicos mais utilizados.

### 1.5.1 Modelo de resposta nominal

O modelo de resposta nominal (Bock [15]) baseia-se no ML2 e surgiu com o objectivo de maximizar a precisão do factor latente estimado, tendo em conta, toda a informação contida nas respostas dos examinandos. Neste modelo, consideram-se as respostas aos itens com mais de duas categorias nominais. A probabilidade de um examinando $j$ seleccionar uma categoria $f$ (de $m_i$ possíveis) do item $i$, é representada por:

$$P_{i,f}(\theta_j) = \frac{e^{Da_{i,f}(\theta_j - b_{i,f})}}{\sum_{h=1}^{m_i} e^{Da_{i,h}(\theta_j - b_{i,h})}} \tag{1.5.1}$$

com $i = 1, 2, ..., I$, $j = 1, 2, ..., J$ e $f = 1, 2, ..., m_i$.

Em cada $\theta_j$, a soma das probabilidades sobre as $m_i$ categorias é 1 (i.e. $\sum_{f=1}^{m_i} P_{i,f}(\theta_j) = 1$). $a_{i,f}$ e $b_{i,f}$ são os parâmetros do item $i$ associados à $f$-ésima categoria.

### 1.5.2 Modelo de resposta gradual

O modelo de resposta gradual foi desenvolvido por Samejima [101] e surge também como uma extensão do ML2. Neste modelo supõe-se que as pontuações das categorias de um item $i$ estão ordenadas da menor para a maior e se denotam por $f = 0, 1, ..., m_i$ onde $m_i + 1$ é o número de categorias do $i$-ésimo item. A probabilidade de um examinando $j$ responder a uma categoria maior ou igual a $f$ no item $i$ é dada por:

$$P^+_{i,f}(\theta_j) = \frac{e^{Da_i(\theta_j - b_{i,f})}}{1 + e^{Da_i(\theta_j - b_{i,f})}} \tag{1.5.2}$$

com $b_{i,f}$ é o parâmetro de dificuldade da $f$-ésima categoria do item $i$.

Os restantes parâmetros do modelo são análogos aos já definidos anteriormente.

Por definição do modelo temos:

$$b_{i,0} \leq b_{i,1} \leq ... \leq b_{i,m_i},$$

uma ordenação entre os níveis de dificuldade das categorias de um dado item.

A probabilidade de um examinando $j$ ter uma pontuação $f$ no item $i$ é dada por:

$$P_{i,f}(\theta_j) = P^+_{i,f}(\theta_j) - P^+_{i,f+1}(\theta_j).$$

Samejima [101] define $P^+_{i,0}(\theta_j) = 1$ e $P^+_{i,m_i+1}(\theta_j) = 0$. Então temos que, na forma logística, o modelo de resposta gradual é o que se apresenta seguidamente:

$$P_{i,f}(\theta_j) = \frac{e^{Da_i(\theta_j - b_{i,f})}}{1 + e^{Da_i(\theta_j - b_{i,f})}} - \frac{e^{Da_i(\theta_j - b_{i,f+1})}}{1 + e^{Da_i(\theta_j - b_{i,f+1})}}. \tag{1.5.3}$$

### 1.5.3 Modelo de escala gradual

Este modelo é um caso particular do modelo de resposta gradual em que se supõe que as pontuações das categorias são igualmente espaçadas, como nas escalas de Likert, mantendo o mesmo número de categorias de respostas para todos os itens. Este modelo foi desenvolvido por Andrich [6] e a sua formulação é a seguinte:

$$P_{i,f}(\theta_j) = \frac{e^{Da_i(\theta_j - b_i + d_f)}}{1 + e^{Da_i(\theta_j - b_i + d_f)}} - \frac{e^{Da_i(\theta_j - b_i + d_{f+1})}}{1 + e^{Da_i(\theta_j - b_i + d_{f+1})}} \tag{1.5.4}$$

com $f = 0, 1, ..., m$, onde:

$b_i$ é o parâmetro de localização do item $i$;

$d_f$ é o parâmetro de categoria.

A fórmula anterior pode ser definida de modo que: $b_i - d_f = b_{i,f}$.

Como $P^+_{i,f}(\theta_j) - P^+_{i,f+1}(\theta_j) \geq 0$, então $d_f - d_{f+1} \geq 0$. Ou seja, devemos ter:

$$d_1 \geq d_2 \geq ... \geq d_m.$$

De notar que os parâmetros da categoria $d_f$ não dependem do item, ou seja, são comuns a todos os itens do teste.

### 1.5.4 Modelo de crédito parcial

O modelo de crédito parcial foi desenvolvido por Masters [82] e é utilizado com os mesmos objectivos do modelo de resposta gradual. É uma extensão do ML1 e, assim, pressupõe que todos os itens tenham o mesmo poder de discriminação. O número de categorias pode variar de item a item no teste. Deste modo, supõe-se que o item $i$ tem $(m_i + 1)$ categorias de resposta ordenadas $(f = 0, 1, ..., m_i)$. A probabilidade de um examinando com factor latente $\theta_j$ obter pontuação na categoria $f$ do item $i$ é dada por:

$$P_{i,f}(\theta_j) = \frac{e^{\sum_{u=0}^{f}(\theta_j - b_{i,u})}}{\sum_{u=0}^{m_i} e^{\sum_{v=0}^{u}(\theta_j - b_{i,v})}} \tag{1.5.5}$$

onde:

$b_{i,f}$ é o parâmetro de item que regula a probabilidade de obter pontuação na categoria $f$ em vez da categoria anterior adjacente $(f-1)$ no item $i$. Cada parâmetro $b_{i,f}$ corresponde ao valor do factor latente em que o examinando tem a mesma probabilidade de responder à categoria $f$ e à categoria $(f-1)$, isto é, $P_{i,f}(\theta_j) = P_{i,f-1}(\theta_j)$. Em geral, define-se $b_{i,0} \equiv 0$.

### 1.5.5 Modelo de crédito parcial generalizado

O modelo de crédito parcial generalizado foi formulado por Muraki [88] e foi baseado no modelo de crédito parcial, considerando que o poder de discriminação não é uniforme para todos os itens. Possui os mesmos pressupostos do ML2 e a sua formulação matemática é a que se apresenta seguidamente:

$$P_{i,f}(\theta_j) = \frac{e^{\sum_{u=0}^{f} Da_i(\theta_j - b_{i,u})}}{\sum_{u=0}^{m_i} e^{\sum_{v=0}^{u} Da_i(\theta_j - b_{i,v})}} \tag{1.5.6}$$

onde as variáveis são definidas como na subsecção anterior.

Se o número de categorias de respostas é $(m_i + 1)$ são estimados $m_i$ parâmetros de categoria do item. O parâmetro $b_{i,f}$ tem o mesmo significado que no modelo de crédito parcial. Adicionalmente, do mesmo modo que no modelo de escala gradual, o parâmetro $b_{i,f}$ pode ser decomposto por $b_{i,f} = b_i - d_f$. Contudo, neste modelo, os valores de $d_f$ não são necessariamente ordenados sequencialmente dentro de um item. O parâmetro $d_f$ é interpretado como a dificuldade relativa da categoria $f$ em comparação com as outras categorias do item.

Na figura 1.5 apresenta-se um exemplo de uma curva característica de um item com 3 categorias de resposta.

A curva característica deste item é constituída por 3 curvas referentes às categorias de resposta do item. Observa-se que a categoria de resposta 1 tem maior probabilidade para níveis do factor latente mais baixos, a categoria 2 apresenta maior probabilidade para níveis do factor latente intermédios, e a categoria 3 apresenta

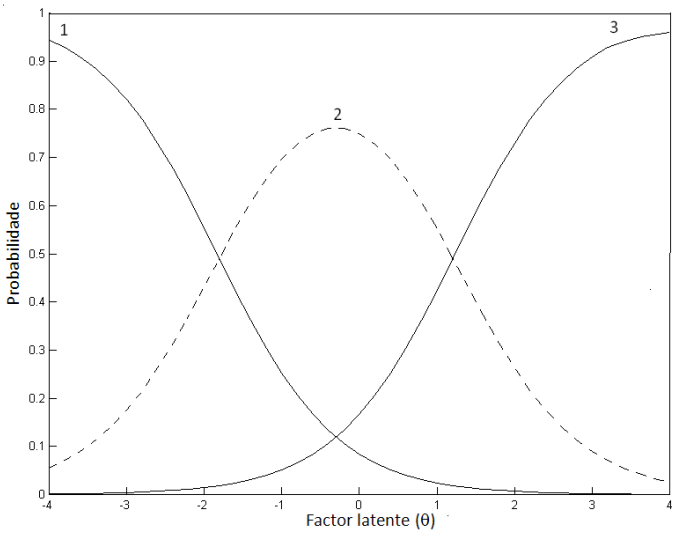



Figura 1.5: Curva característica de um item com 3 categorias de resposta

maior probabilidade para níveis do factor latente mais elevados.

## 1.6 Modelos para grupos múltiplos

Frequentemente, existe interesse em incorporar modelos que permitam estudar examinandos que provêm de diferentes grupos ou populações, dado que apresentam características diferentes, próprias de cada um desses grupos. Nesse sentido, surgiram os MRI, que têm em consideração as características particulares desses grupos, entre os quais o denominado modelo para grupos múltiplos, desenvolvido por Bock e Zimowski [20].

O modelo proposto Bock e Zimowski [20] é uma generalização dos modelos logísticos para o caso de dois ou mais grupos. Nesse sentido, para o caso da generalização do ML3, a probabilidade de um examinando $j$ do grupo $l$, com factor latente $\theta_{jl}$, responder correctamente ao item $i$ é dada por:

$$P(U_{ijl} = 1|\theta_{jl}) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_{jl}-b_i)}}{1 + e^{Da_i(\theta_{jl}-b_i)}} \tag{1.6.1}$$

com $i = 1, 2, ..., I$, $j = 1, 2, ..., J_l$ e $l = 1, ..., L$, onde:

$U_{ijl}$ é uma variável dicotómica que assume o valor 1, quando o examinando $j$ do grupo $l$ responde correctamente ao item $i$, ou o valor 0, quando o examinando não responde correctamente ao item; $\theta_{jl}$ representa o factor latente do $j$-ésimo examinando do grupo $l$. Os restantes parâmetros já foram descritos anteriormente.

Bock e Zimowski [20] referem que o modelo para grupos múltiplos apresenta grande aplicabilidade no funcionamento diferencial do item, quando existe "desgaste"dos parâmetros dos itens, na equalização de grupos não-equivalentes e na equalização vertical.

## 1.7 Função de informação do item

A utilização de MRI permite o uso de uma medida bastante utilizada complementarmente com a CCI que é a função de informação do item. Ela permite quantificar a informação com que o item contribui para a medida do factor latente. A função de informação de um item é dada por:

$$I_i(\theta) = \sum_{f=1}^{m_i} \frac{\left[\frac{d}{d\theta_j} P_{i,f}(\theta_j)\right]^2}{P_{i,f}(\theta_j)} - \frac{d^2}{d\theta_j^2} P_{i,f}(\theta_j) \tag{1.7.1}$$

onde,

$I_i(\theta)$ é a "informação" fornecida pelo item $i$ na escala do factor latente $\theta$;

$P_{i,f}(\theta_j)$ é a probabilidade de um examinando com factor latente $\theta_j$ obter pontuação na categoria $f$ do item $i$.

Na expressão anterior, considera-se que os itens são politómicos. Em particular, no caso de itens dicotómicos, a função de informação do item é a seguinte:

$$I_i(\theta) = \frac{\left[\frac{d}{d\theta_j} P(U_{ij} = 1|\theta_j)\right]^2}{P(U_{ij} = 1|\theta_j)(1 - P(U_{ij} = 1|\theta_j))} \tag{1.7.2}$$

onde,

$P(U_{ij} = 1|\theta_j)$ é a probabilidade de acerto no item dada por um modelo dicotómico.

## 1.8 Função de informação do teste

A informação fornecida pelo teste foi desenvolvida por Birnbaum [14], é dada pela soma das funções de informação dos itens em $\theta$ e denota-se por $I(\theta)$:

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta) \tag{1.8.1}$$

Note-se que os itens contribuem, independentemente, para a função de informação do teste, mas a contribuição individual de cada item pode ser determinada sem conhecimento dos outros itens do teste.

A importância da informação que um teste fornece está inversamente relacionada com o erro-padrão de estimação:

$$EP(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{1.8.2}$$

É importante realçar que essas medidas de informação dependem do valor de $\theta$.

A figura 1.6 ilustra como a função de informação dos itens constitui a função de informação do teste.

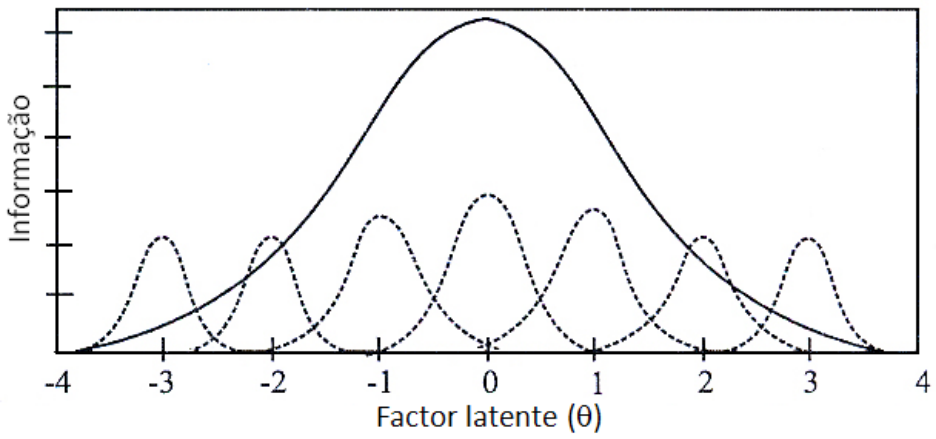



Figura 1.6: Funções de informação de 7 itens (curvas a tracejado) e do teste

A função de informação do item e função de informação do teste são importantes para a selecção de itens e na construção de testes, elaborados para um determinado nível do factor latente dos examinandos.

## 1.9 Equalização e *linking*

### 1.9.1 Introdução

Nas avaliações das classificações de diferentes testes que aferem o mesmo factor latente, ou até mesmo factores latentes diferentes, é necessário estabelecer a comparabilidade das classificações. Existem técnicas cujo objectivo fundamental consiste em justificar e operacionalizar funções de ligação entre dois testes diferentes. Há dois procedimentos fundamentais de comparação de classificações: equalização e *linking*.

Equalização define-se como o método estatístico usado para ajustar a classificação dos examinandos a uma escala única, de modo a que subtestes diferentes sejam aplicados a amostras diferentes de examinandos, eventualmente em momentos diferentes no tempo, e, ainda assim, a classificação obtida seja passível de comparação (Kolen e Brennan [71]). A equalização é utilizada quando os testes ou formas diferentes de um teste sejam construídos de acordo com dois requisitos: aferem o mesmo factor latente e possuem as mesmas especificações estatísticas (tais como: dificuldade, formato e tamanho de teste, entre outras). Há dois tipos de equalização: horizontal e vertical. A equalização horizontal consiste na comparação das competências desenvolvidas em diferentes populações de examinandos, que estejam, por exemplo no mesmo nível de ensino, e a equalização vertical permite comparar as competências desenvolvidas ao longo do tempo, por exemplo, ao longo da trajectória escolar dos examinandos.

*Linking* refere-se ao procedimento estatístico que visa relacionar as classificações de testes ou formas de teste com diferentes especificações (que diferem em conteúdo e/ou nível de dificuldade) e/ou no caso em que aferem diferentes factores latentes. Consideram-se três tipos de *linking*: calibração, projecção e moderação.

i) Calibração: os testes são construídos de acordo com a mesma estrutura ou matriz (mesmo conteúdo), mas com precisão e dificuldade diferentes (especificações estatísticas diferentes). Neste caso, a comparação das classificações é

feita de tal forma que a classificação num teste se torna igual à classificação do outro, sendo necessário mais do que uma função estatística para efectuar a comparação;

ii) Projecção: as tarefas nos dois testes são diferentes, bem como as condições de aplicação dos mesmos ou, ainda, os dois testes aferem factores latentes diferentes. Utilizam-se procedimentos estatísticos, como a regressão para obter características do teste $M$ com base nas informações obtidas no teste $N$. Neste caso, existe algum risco ao efectuar a comparação entre dois testes, uma vez que as classificações podem depender de uma série de outros factores que são diferentes nos examinandos que realizaram testes diferentes. Por exemplo, usar as classificações em Português para prever os resultados em Matemática;

iii) Moderação: assume-se que dois testes medem factores latentes diferentes, mas mesmo assim pretende-se comparar as classificações dos dois testes. Normalmente, são utilizadas as mesmas técnicas para comparação de classificações nos dois testes, mas as relações encontradas não têm suporte para interpretar essas relações quando os testes e os examinandos são totalmente diferentes.

Os problemas principais do *linking* consistem em determinar relações existentes entre duas medidas diferentes de um factor latente e interpretar correctamente esssas relações.

### 1.9.2 Revisão da literatura

Segundo Dorans [39], em 1944, Ledyard Tucker foi considerado o mentor de muitos desenvolvimentos teóricos dos procedimentos de equalização, nomeadamente, do ajuste de escalas via itens comuns. Angoff [7] designou este procedimento por equalização de Tucker. Flanagan [49] descreveu algumas técnicas de equalização e discutiu alguns temas e problemas relacionados com este assunto. Angoff [7] descreveu procedimentos utilizados na equalização de testes, nomeadamente, de equa-

lização linear e de equalização equipercentílica. Lord [77] apresentou os requisitos de equalização nos MRI. Petersen *et al.* [94] apresentaram vários procedimentos empíricos utilizados na equalização. Dorans [38] desenvolveu a equalização linear de Levine baseada no pressuposto da invariância do erro padrão de medida. Harris e Crouse [61] definiram quatro requisitos para efectuar a equalização, em particular, a invariância da população. Kolen e Brennan [70] distinguiram os dois tipos de equalização, horizontal e vertical. Desde 1995, têm sido publicados muitos trabalhos de investigação relacionados com estimação de erros padrão de equalização e a equalização de MRI politómicos. Kolen e Brennan [71] publicaram os princípios dos procedimentos de equalização e de outras metodologias similares, como *scaling* e *linking*. Estes autores descreveram as propriedades, os pressupostos estatísticos e os procedimentos computacionais dos diversos tipos de equalização, tanto clássicos, como recorrendo a MRI. Von Davier, Holland e Thayer [32] introduziram o método de Kernel na equalização de testes. Holland, Dorans e Petersen [63] descreveram os procedimentos de equalização mais utilizados actualmente e desenvolveram mais aprofundadamente a metodologia *linking*.

No que se refere às aplicações, a equalização de testes nos anos 80 teve um grande desenvolvimento, que se deveu essencialmente ao aumento do número, variedade e forma dos testes, à diversidade de classificações obtidas nos diferentes testes e ao aumento dos trabalhos publicados na área da educação. Em 1978, no *Program Statistics Research Project* do *Educational Testing Service* (ETS), foi criado um novo grupo de investigação estatística (*Research Statistics Group*), com vista à investigação de problemas relevantes que surgiram nos programas de testes do ETS. Paul Holland foi o responsável pela investigação sobre a equalização dos testes e Rubin [100] trabalhou na validação dos seus pressupostos. Os trabalhos desenvolvidos no *Program Statistics Research Project* vieram dar origem a uma conferência sobre equalização de testes em 1980, cujos trabalhos foram publicados em Holland e Rubin [64]. Holland e Thayer [65] desenvolveram um novo método de equaliza-

ção designado *Pré-Equating*. Em 1999, os relatórios, *Uncommon Measures* de Feuer *et al.* [48] e *Embedding Questions* de Koretz, Bertenthal e Green produzidos pelo *National Research Council*, apresentaram vários exemplos de equalização em testes educacionais. Livingston [75] publicou um manual para investigadores da área da equalização, sem recurso aos MRI, onde descreveu muitos procedimentos que se utilizavam na prática. Têm sido publicadas diversas aplicações dos procedimentos de equalização, como: Ghadan [53], que estuda os efeitos do tamanho da amostra na equalização e Dorans [40], Guilera e Gómez [54] e Chen *et al.* [23], que recorreram à utilização MRI para realizar equalização e *linking* de testes na área das Ciências da Saúde.

### 1.9.3 Requisitos

Segundo von Davier, Holland e Thayer [32], para se utilizar qualquer método de equalização/*linking* devem ser verificados os seguintes requisitos:

1) igualdade dos factores latentes: não devem ser equalizados testes que afiram diferentes factores latentes;

2) igualdade da fiabilidade: não devem ser equalizados testes que afiram o mesmo factor latente mas que difiram na fiabilidade;

3) simetria: a função de equalização utilizada para equalizar as classificações do teste $M$ em ordem às do teste $N$ deve ser a inversa da função de equalização usada para equalizar as classificações do teste $N$ em ordem às do teste $M$;

4) equidade: a classificação de um examinando deve ser independente do teste utilizado para a equalizar;

5) invariância da população: a função de equalização utilizada para aferir as classificações nos testes $M$ e $N$ não deve depender da amostra considerada,

isto é, a função de equalização usada para estabelecer a ligação entre as classificações dos testes $M$ e $N$ deve ser invariante à população.

Kolen e Brennan [71] referem que para efectuar o equalização e/ou *linking* de testes é necessário:

1) definir o objectivo da comparação;

2) executar o processo de desenvolvimento de um conjunto de itens, isto é, construir os testes;

3) projectar e implementar o plano da recolha de dados;

4) escolher um ou mais procedimentos estatísticos para efectuar a comparação dos testes;

5) analisar e avaliar os resultados;

6) verificar as condições de padronização e os procedimentos de controle/ajuste dos dados.

Os procedimentos para o desenvolvimento dos testes, os planos de recolha de dados e os procedimentos estatísticos utilizados no *linking* são basicamente os mesmos dos que os utilizados na equalização. Nesse sentido, daqui em diante referimo-nos ao *linking* e indicamos as principais diferenças comparativamente com a equalização.

### 1.9.4 Desenvolvimento dos testes

No procedimento *linking*, Feuer *et al.* [48] consideram três fases no desenvolvimento dos testes:

1) Definição da estrutura: delineação do objectivo e da extensão (em termos de conteúdos específicos das áreas e factores latentes) do domínio que vai ser representado na avaliação;

2) Especificações do teste: combinação de conteúdos por áreas específicas e dos formatos dos itens, número de itens, critérios de avaliação e outros;

3) Selecção dos itens: os itens são seleccionados de forma a representarem as especificações definidas.

### 1.9.5 Planos de recolha de dados

O *linking* depende do planeamento da recolha de dados, havendo três planos fundamentais (figura 1.7):

i) uma amostra aleatória de uma população: a amostra é repartida em tantos subgrupos quanto o número de testes. Este modelo é raramente utilizado porque apresenta algumas limitações: o cansaço, devido ao uso de testes excessivamente longos e complicados e a possível influência da ordem dos itens nas classificações finais. Para evitar o efeito de cansaço no teste, este é dividido em dois subtestes, $U$ e $V$, e são considerados dois subgrupos de examinandos. A um deles é administrado o subteste $U$, seguido do subteste $V$ e vice-versa para o outro subgrupo. Os subtestes podem ser aplicados na mesma ocasião ou em ocasiões diferentes. Para se trabalhar com este tipo de recolha de dados, o número de examinandos deve ser razoavelmente grande, em geral, uma amostra de 500 examinandos.

ii) duas amostras aleatórias da mesma população: consideram-se os subtestes $U$ e $V$ do teste em que uma amostra responde ao subteste $U$ e a outra amostra responde ao subteste $V$. Uma vantagem deste tipo de recolha de dados é que cada amostra responde somente a um subteste, reduzindo o tempo e o cansaço na aplicação.

iii) duas amostras aleatórias de populações diferentes com itens âncora: dois testes $M$ e $N$ são aplicados a duas amostras de examinandos na mesma ocasião,

mas em ambos os testes há um número $X$ de itens comuns, denominados itens âncora. Kolen e Brennan [71] referem que o número de itens âncora entre os testes deve representar pelo menos 20% dos itens que vão ser equalizados. Este é o método mais utilizado e é considerado o mais adequado para efectuar o *linking*.

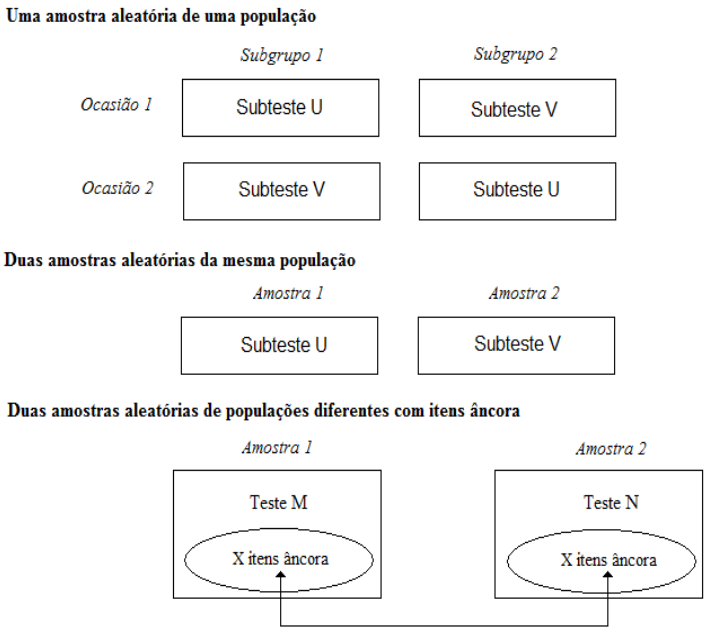



Figura 1.7: Planos de recolha de dados

### 1.9.6 Métodos/Procedimentos

Considerando os diferentes planos da recolha de dados, mencionados na subsecção 1.9.5, o *linking* pode ser efectuado via população ou via itens comuns.

No plano de recolha de dados i), o *linking* é feito via população, pela aplicação de um dos procedimentos seguintes:

I - efectua-se a estimação conjunta dos parâmetros dos itens e do factor latente em ambos os subtestes $U$ e $V$;

II - obtêm-se as estimativas separadamente, desde que a distribuição do factor latente tenha os mesmos parâmetros na estimação em $U$ e em $V$.

No plano ii) da recolha de dados, o *linking* é feito via população, pela aplicação do procedimento II ou pela utilização de um dos métodos abaixo descritos.

No plano da recolha de dados iii), o *linking* é feito via itens comuns, pela aplicação de um dos métodos descritos seguidamente.

Kolen e Brennan [71] consideram quatro métodos para efectuar o *linking*[2]: linear, média, paralelo linear e equipercentílico.

No método linear, postula-se que as estimativas do factor latente obtidas no teste $M$ e no teste $N$ estão linearmente relacionadas. Considera-se que os parâmetros $\theta$, $a_i$ e $b_i$ são estimados separadamente para os testes $M$ e $N$, obtendo-se respectivamente $\hat{\theta}_N$, $\hat{\theta}_M$, $(\hat{a}_N, \hat{b}_N)_i$, $(\hat{a}_M, \hat{b}_M)_i$. Por simplicidade, omite-se o índice $i$ referente ao item. A relação linear entre as duas escalas do factor latente pode ser expressa do seguinte modo:

$$\hat{\theta}_N = A\hat{\theta}_M + B \tag{1.9.1}$$

e os parâmetros dos itens do modelo ML3 relacionam-se da seguinte forma:

$$\hat{a}_N = \frac{\hat{a}_M}{A} \tag{1.9.2}$$

$$\hat{b}_N = A\hat{b}_M + B \tag{1.9.3}$$

$$\hat{c}_N = \hat{c}_M \tag{1.9.4}$$

onde $A$ e $B$ são constantes.

O método linear assenta fundamentalmente na determinação das constantes $A$ e $B$. Para tal, a partir da equação 1.9.5:

[2]Apenas no método Paralelo Linear não é utilizado no procedimento equalização.

$$\frac{\hat{\theta}_N - \mu_N}{\sigma_N} = \frac{\hat{\theta}_M - \mu_M}{\sigma_M} \tag{1.9.5}$$

obtém-se:

$$A = \frac{\sigma(\hat{b}_N)}{\sigma(\hat{b}_M)} \tag{1.9.6}$$

$$A = \frac{\mu(\hat{a}_M)}{\mu(\hat{a}_N)} \tag{1.9.7}$$

$$A = \frac{\sigma(\hat{\theta}_N)}{\sigma(\hat{\theta}_M)} \tag{1.9.8}$$

$$B = \mu(\hat{b}_N) - A\mu(\hat{b}_M) \tag{1.9.9}$$

$$B = \mu(\hat{\theta}_N) - A\mu(\hat{\theta}_M) \tag{1.9.10}$$

onde

$\sigma(\hat{b}_N)$ e $\sigma(\hat{b}_M)$ são os desvios padrão das estimativas do parâmetro de dificuldade dos itens âncora no teste $N$ e no teste $M$, respectivamente;

$\mu(\hat{a}_M)$ e $\mu(\hat{a}_N)$ são as médias das estimativas do parâmetro de discriminação dos itens âncora nos testes $M$ e $N$, respectivamente;

$\sigma(\hat{\theta}_N)$ e $\sigma(\hat{\theta}_M)$ são os desvios padrão das estimativas do factor latente do teste $N$ e do teste $M$, respectivamente;

$\mu(\hat{b}_N)$ e $\mu(\hat{b}_M)$ são as médias das estimativas do parâmetro de dificuldade dos itens âncora nos testes $N$ e $M$, respectivamente;

$\mu(\hat{\theta}_N)$ e $\mu(\hat{\theta}_M)$ são as médias das estimativas do factor latente no teste $N$ e no teste $M$, respectivamente.

Segundo Kolen e Brennan [71], quando são usados os valores obtidos em (1.9.6) e (1.9.9) nas expressões (1.9.1), (1.9.2) e (1.9.3), o procedimento é usualmente conhecido por Média-Desvio (Marco [81]) e quando são utilizados os valores obtidos em

(1.9.7) e (1.9.9) nas expressões (1.9.1), (1.9.2) e (1.9.3), o procedimento é designado por Média-Média (Loyd e Hoover [79]).

Com vista à comparação dos diferentes procedimentos de equalização linear, Kolen e Brennan [71] referem dois critérios baseados nas Curvas Características dos Itens e denominados *Hdiff* e *SLdiff*. Segundo estes autores, no critério *Hdiff*, Haebara [56] utiliza uma função para expressar a diferença entre as curvas características dos itens nos testes $M$ e $N$ como sendo a soma do quadrado da diferença das Curvas Características de cada Item, para examinandos com um determinado factor latente. Para um dado $\hat{\theta}_j$, a referida soma para os itens âncora ($x = 1, 2, ..., X$), pode ser escrita como:

$$Hdiff(\hat{\theta}_j) = \sum_{x=1}^{X} [P_x(\hat{\theta}_{Nj}) - P_x(\hat{\theta}_{Mj})]^2 \tag{1.9.11}$$

No critério *SLdiff*, Stocking e Lord [108], utilizaram o quadrado da diferença entre as curvas características para um dado $\hat{\theta}_j$.

$$SLdiff(\hat{\theta}_j) = [\sum_{x=1}^{X} P_x(\hat{\theta}_{Nj}) - \sum_{x=1}^{X} P_x(\hat{\theta}_{Mj})]^2 \tag{1.9.12}$$

Kolen e Brennan [71] referem que, para ambos os critérios, quanto menor for o valor obtido "melhor" é o procedimento utilizado e que estes critérios devem ser calculados para vários valores de $\hat{\theta}_j$.

No método média postula-se que as estimativas do factor latente nos dois testes, que se situam a distâncias iguais das respectivas médias, podem ser igualadas. É um caso particular do método linear no qual $A = 1$, porém não constitui a melhor escolha, apesar da sua simplicidade, a não ser que as amostras sejam pequenas.

O método paralelo linear (Dorans e Holland [41]) é utilizado para grupos múltiplos. A única diferença estatística entre este e o método linear é que o desvio para os subgrupos é dividido pelo desvio padrão do grupo combinado.

O método equipercentílico postula que as estimativas do factor latente em dois testes são equivalentes, se correspondem ao mesmo percentil. Nesse sentido, as

diferenças na dificuldade entre os testes são descritas por uma transformação não linear do seguinte modo:

$$e_N = G^{-1}F_M \tag{1.9.13}$$

onde

$e_N$ é a função percentílica utilizada para converter as estimativas do factor latente do teste $M$ para o teste $N$;

$G^{-1}$ é a inversa da função de distribuição cumulativa da função $e_N$, sendo $G$ a função de distribuição cumulativa da população no teste $N$;

$F_M$ é a função de distribuição cumulativa das estimativas do factor latente de $M$.

Este método apresenta inúmeras vantagens relativamente aos métodos linear, média e paralelo linear, nomeadamente: as comparações equipercentílicas pertencem ao intervalo das classificações, o que pode não acontecer nos outros métodos; as relações estabelecidas entre os testes não são assumidas como lineares; a função de distribuição cumulativa das classificações transformadas do teste $M$ é, aproximadamente, dada pela função de distribuição cumulativa de $N$; os momentos das classificações transformadas do teste $M$ (isto é, média, variância, assimetria e achatamento) são, aproximadamente, as mesmas que as de $N$.

São ainda referenciados na literatura (Andrade, Tavares e Valle [5]) outros dois métodos de equalização via itens comuns: calibração simultânea e calibração com parâmetros de itens fixos.

No que se refere ao procedimento de equalização via itens comuns designado calibração simultânea, os itens são calibrados[3] utilizando as respostas dos examinandos em ambos os grupos simultaneamente. Este tipo de equalização é feito durante o processo de calibração, pelo uso do modelo de resposta ao item para grupos múltiplos (equação 1.6.1), em que são os itens comuns que fazem a ligação entre as populações envolvidas. Para obter as estimativas dos parâmetros do modelo, pode-se recorrer

[3]Calibração é o processo de estimação dos parâmetros dos itens.

à utilização de procedimentos de estimação de máxima verosimilhança ou a procedimentos de estimação bayesianos[4]. O software mais utilizado pelos investigadores para obter as estimativas dos parâmetros do modelo é o Bilog-Mg (Zimowski *et al.* [118]). A utilização deste software permite que a equalização seja feita automaticamente no próprio processo de estimação e, consequentemente, todos os parâmetros estejam na mesma escala, o que possibilita comparações e a construção de escalas de conhecimento interpretáveis. Nos procedimentos de estimação de modelos que envolvem mais do que uma população, existem problemas de indeterminação da escala. Para resolver estes problemas, o procedimento que se deve adoptar, é considerar uma das populações como sendo a referência, que é definida como tendo média zero e desvio padrão um, e, consequentemente, as restantes populações serão posicionadas em relação à população de referência. O procedimento de equalização simultânea é considerado o melhor exemplo do uso e importância da equalização e ilustra o maior avanço dos MRI, comparativamente com a TCT. As vantagens apontadas para o uso deste procedimento de equalização são as que se apresentam seguidamente. Como a equalização é feita automaticamente no próprio processo de estimação, não existem diferenças nas estimativas dos parâmetros devidas ao procedimento de equalização escolhido. Em particular, segundo Andrade, Tavares e Valle [5], na presença de um número de populações superior ou igual a 5, ao aplicar-se algum procedimento de equalização entre os testes existem erros (por exemplo, relativos à regressão) que estão associados a cada equalização entre duas populações, o que faz com que se acumulem erros ao longo dos procedimentos de estimação seguintes. Outra das vantagens para a utilização da calibração simultânea é a exigência de um menor número de itens comuns, comparativamente com outros tipos de equalização via itens comuns, para produzir resultados similares.

O procedimento de equalização denominado itens comuns de calibração com parâmetros de itens fixos, usa-se quando se pretende calibrar determinados itens e

[4]Mais detalhes sobre estes procedimentos de estimação podem ser encontrados em Andrade, Tavares e Valle [5], capítulo 5.

manter os parâmetros de outros que já foram calibrados. Para fixar os parâmetros de alguns itens e calibrar os restantes, utiliza-se o software Bilog-Mg (Zimowski *et al.* [118]). Neste caso, utiliza-se um procedimento de estimação bayesiano, em que para fixar os parâmetros dos itens já calibrados, se utilizam distribuições *a priori* convenientes para os parâmetros desses itens, cujas médias são os próprios valores dos parâmetros que pretendemos fixar e os desvios padrão são tão pequenos que a distribuição se torna praticamente degenerada naquele ponto. Na prática, todos os parâmetros são estimados novamente, mas a convergência dos itens já calibrados é, artificialmente, induzida para os valores que se pretendem obter. Outra forma de fixar os parâmetros dos itens já calibrados, é definir, convenientemente, os valores para esses parâmetros. Este segundo procedimento pode proporcionar alguns problemas de convergência, no caso de existir mais do que um grupo de examinados envolvido. Assim, quando há mais do que uma população, o procedimento que se deve adoptar, é a utilização do software Bilog-Mg (Zimowski *et al.* [118]), para fixar os itens calibrados e estimar os restantes itens. Por vezes a aplicação do último procedimento descrito apresenta problemas com a métrica, pelo facto de existir mais do que uma população envolvida. Assim, para resolver os problemas de indeterminação de escala, define-se o grupo de referência do mesmo modo que foi referido no procedimento de calibração simultânea. O objectivo é que os itens já calibrados tenham os parâmetros fixados na escala da população de referência e, assim, equalizam-se os novos parâmetros dos itens de acordo com a escala da população de referência.

## 1.10 Considerações gerais

Neste capítulo, apresentámos os principais modelos de resposta ao item unidimensionais existentes para descrever a relação entre os examinandos e os itens de instrumentos. No capítulo 3 apresentam-se diversas aplicações, usando dados reais, para esta classe de modelos. Os modelos unidimensionais têm inúmeras vantagens, mas nem sempre os examinandos possuem apenas um factor latente quando respon-

dem a um item do instrumento e os itens podem requerer mais do que um factor latente para serem respondidos correctamente. Assim, os modelos unidimensionais são convenientes em certas condições, mas, por vezes, existe a necessidade do uso de modelos de resposta ao item mais complexos que permitam estabelecer, com maior precisão, as relações entre os examinandos e as características dos itens do instrumento. No capítulo seguinte, apresentam-se extensões de modelos unidimensionais de resposta ao item que consideram múltiplos parâmetros para os examinandos e que são designados por modelos multidimensionais.

# Capítulo 2

# Modelos Multidimensionais de Resposta ao Item

## 2.1 Introdução

Nos Modelos Multidimensionais de Resposta ao item (MMRI), as respostas aos itens estão associadas a múltiplos factores latentes. Nos MMRI consideram-se múltiplos parâmetros para os factores latentes do examinando e um vector de parâmetros que caracterizam os itens. Esta classe de modelos permite avaliar o comportamento dos examinandos, isto é, a resposta ao item, dados os factores latentes dos examinandos e os parâmetros que representam as características do item. A utilização desta classe de modelos é importante para reflectir o nível de factor latente/habilidade nas diferentes dimensões para cada examinando.

O desenvolvimento de modelos multidimensionais de resposta ao item iniciou-se com os trabalhos de Rasch [96]. Este autor apresentou uma generalização do seu modelo unidimensional que incluia a possibilidade do factor latente do examinando ser representado por um vector em vez de um escalar. Contudo, apesar das características multidimensionais deste modelo, ainda se evidenciava o desafio de obter as estimativas dos elementos do vector de factores latentes, mantendo as propriedades

do modelo de Rasch. O modelo de Rasch multidimensional foi pouco utilizado, devido à complexidade dos procedimentos e pelos problemas que se verificavam na especificação das funções que representavam os factores latentes dos examinandos.

Os requisitos principais de um modelo multidimensional foram expostos no 16º capítulo do livro de Lord e Novick [78]. Este capítulo inclui as definições de espaço latente completo[1] e do pressuposto da independência local e é composto, essencialmente, pela discussão do significado dos parâmetros dos itens e do seu uso na resolução de problemas práticos. Adicionalmente, Lord e Novick [78] mostraram a relação existente entre o modelo de resposta ao item unidimensional de ogiva normal e o modelo de análise factorial comum.

Samejima [102] desenvolveu um MMRI para itens que apresentam resposta contínua. Apesar do modelo desenvolvido por Samejima ter sido formalmente o primeiro a ser apresentado como MMRI, com excepção do trabalho de Bejar [13], verifica-se que este modelo não foi praticamente aplicado. A justificação para a sua falta de aplicação é, provavelmente, o facto de não ser frequente que em testes psicológicos e educacionais exista este tipo de resposta no item.

Neste capítulo, apresentamos a especificação formal dos principais modelos multidimensionais de resposta ao item, expomos os procedimentos de estimação actualmente existentes para a estimação dos parâmetros dos MRI e propomos um algoritmo que usa abordagem bayesiana e MCMC para estimar os parâmetros dos itens e dos factores latentes do MMRI logístico de 2 parâmetros.

## 2.2 Modelos multidimensionais

O tipo de MMRI distingue-se pela forma como é estabelecida a relação entre as coordenadas de um vector $\theta$ com as características do item, para especificar a probabilidade de resposta ao item. Existem dois tipos principais de MMRI: compen-

[1] É definido pelo vector $\theta$ com $k$ componentes.

satórios e não compensatórios (Reckase [97]).

Os modelos compensatórios baseiam-se numa combinação linear das coordenadas de $\theta$. Para especificar a probabilidade de resposta, a combinação linear é usada na forma ogiva normal ou logística. A combinação linear das coordenadas $\theta$ pode fornecer a mesma soma com várias combinações dos valores de $\theta$. Neste tipo de modelos, os factores interagem de forma a que a diminuição da magnitude de um factor latente possa ser compensada pelo aumento na magnitude noutros factores, isto é, se uma é baixa, a soma será a mesma se outra coordenada de $\theta$ é suficientemente alta. Seguidamente apresentam-se os principais modelos multidimensionais compensatórios.

### Modelo multidimensional logístico de 2 parâmetros

O ML2 (equação 1.4.2) tem como expoente $a_i(\theta - b_i)$, o que é idêntico a $a_i\theta - a_ib_i$. Se substituirmos $-a_ib_i$ por $d_i$, a expressão anterior é dada por $a_i\theta + d_i$. Assim, da extensão do ML2 ao modelo multidimensional obtém-se considerando $\theta_j$ um vector $(1 \times k)$ referente ao examinando $j$, $a_i$ é um vector $(1 \times k)$ relativo aos parâmetros de discriminação do item $i$, em que $k$ é o número de dimensões do factor latente e o termo $d_i$ é um escalar que representa a dificuldade de cada item. O modelo multidimensional logístico de 2 parâmetros (MML2) é especificado por (2.2.1):

$$P(U_{ij} = 1|\theta_j, a_i, d_i) = \frac{e^{a_i\theta'_j + d_i}}{1 + e^{a_i\theta'_j + d_i}}. \tag{2.2.1}$$

O expoente de $e$ neste modelo pode ser escrito da seguinte forma:

$$a_i\theta'_j + d_i = a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + ... + a_{ik}\theta_{jk} + d_i = \sum_{l=1}^{k} a_{il}\theta_{jl} + d_i. \tag{2.2.2}$$

O expoente é uma função linear de elementos de $\theta$ com o parâmetro $d$ como a ordenada na origem e os elementos do vector $a$ como os parâmetros de inclinação/dis-

criminação.

### Modelo multidimensional logístico de 3 parâmetros

O modelo multidimensional logístico de 3 parâmetros (MML3) surgiu a partir do MML2, considerando o parâmetro de probabilidade de resposta ao acaso. Nesta classe de modelos, a probabilidade de um examinando com valores baixos de $\theta$ responder correctamente a um item não está relacionada com os factores latentes aferidos pelos itens do teste. Nesse sentido, este modelo contém um escalar $c_i$ para cada item $i$, que corresponde ao parâmetro de probabilidade de acerto ao acaso. É uma extensão multidimensional do ML3 (equação 1.4.3) e a sua formulação matemática é a seguinte:

$$P(U_{ij} = 1|\theta_j, a_i, c_i, d_i) = c_i + (1 - c_i)\frac{e^{a_i\theta'_j+d_i}}{1 + e^{a_i\theta'_j+d_i}}. \tag{2.2.3}$$

Os parâmetros deste modelo já foram definidos anteriormente.

### Modelo multidimensional logístico de 1 parâmetro

O modelo multidimensional logístico de 1 parâmetro (MML1) (Adams, Wilson e Wang [2]), também designado modelo multidimensional de Rasch, é dado por:

$$P(U_{ij} = 1|\theta_j, a_i, d_i) = \frac{e^{a_i\theta'_j+d_i}}{1 + e^{a_i\theta'_j+d_i}}. \tag{2.2.4}$$

onde $a_i$ é especificado *a priori* e é tratado como uma constante. No MML1, $a_i$ é uma característica do item $i$, é especificada pelo especialista que desenvolveu o teste e que, em geral, lhe atribui valores inteiros.

No caso do MML1, a qualidade do ajuste do modelo aos dados vai depender da forma como o construtor do teste especifica os valores do vector $a_i$. Mais detalhes podem ser encontrados em Reckase [97].

### Modelo multidimensional da ogiva normal

O desenvolvimento inicial de modelos multidimensionais foi feito, recorrendo à utilização da função ogiva normal, para representar a relação entre a localização num espaço multidimensional e a probabilidade de resposta correcta a um item do teste. A formulação matemática mais geral do modelo multidimensional de ogiva normal (Bock e Schilling [19]; McDonald [85]; Samejima [102]) considera o parâmetro de dificuldade geral do item, os parâmetros de discriminação específicos de cada dimensão e o parâmetro de probabilidade de acerto ao acaso, e é dada por:

$$P(U_{ij} = 1|\theta_j, a_i, c_i, d_i) = c_i + (1 - c_i)\frac{1}{\sqrt{2\pi}} \int_{-z_i(\theta_j)}^{\infty} e^{\frac{-t^2}{2}} dt \qquad (2.2.5)$$

onde $z_i(\theta_j) = a_i\theta' j + d_i$. Os parâmetros restantes já foram definidos anteriormente.

A forma apresentada neste modelo define a probabilidade de resposta correcta a um item como a área sob a distribuição normal padrão de $-z_i(\theta_j)$ até $\infty$. Em particular, no caso em que $c_i = 0$, obtém-se o modelo multidimensional de 2 parâmetros.

Um exemplo da existência da multidimensionalidade compensatória é um instrumento composto por itens que têm múltiplas estratégias de resolução em que a falta de uma habilidade cognitiva naturalmente compensa as outras.

Nos MMRI não compensatórios pressupõe-se que todos os factores latentes são importantes para se obter a resposta correcta ao item. Neste tipo de modelos considera-se que para o item ser respondido correctamente seja necessário concluir adequadamente múltiplas tarefas, cada qual associada a um dos factores latentes. A probabilidade de resposta correcta ao item é então representada como sendo o produto das probabilidades de se responder independentemente a cada tarefa em cada componente/parte[2].

[2]Mais detalhes podem ser encontrados em Sympson [109].

## 2.3 Procedimentos de estimação

Nos modelos de resposta ao item são conhecidas, em geral, as respostas dos examinandos aos itens e são desconhecidos os parâmetros que caracterizam o item e o factor latente do examinando. Assim, a estimação envolve dois tipos de parâmetros, os factores latentes dos examinandos e os parâmetros dos itens. A estimação do factor latente é, geralmente, feita recorrendo à estimação pela máxima verosimilhança (por exemplo, Baker e Kim [11]), ou a métodos bayesianos (por exemplo, Baker e Kim [11]) como a estimação pela moda da *posteriori* e a estimação pela média da *posteriori*.

O método da máxima verosimilhança consiste em maximizar a função de verosimilhança das respostas ao item do examinando. Uma das desvantagens apontadas para a utilização deste método é o facto de não estar definido para alguns padrões de resposta, isto é, não é possível estimar o factor latente de examinandos que acertaram em todos os itens ou de examinandos que erraram todos os itens. Adicionalmente, os estimadores de máxima verosimilhança são assintóticamente consistentes apenas para testes constituídos por um grande número de itens (Andrade, Tavares e Valle [5]).

A estimação pela moda da *posteriori* usa a distribuição *a priori* do factor latente conjugada com a função de verosimilhança para estimar o factor latente, pela maximização da distribuição *a posteriori*. A principal vantagem deste método é que está definido para qualquer padrão de resposta. Uma das desvantagens da utilização deste método é que depende da distribuição definida *a priori* para o factor latente.

A estimação pela média da *posteriori* permite usar qualquer tipo de padrão de resposta e possui menor erro médio na estimação do factor latente quando as *prioris* são definidas convenientemente. A má escolha das *prioris*, nomeadamente *prioris* com variâncias pequenas e médias distantes dos valores verdadeiros dos parâmetros, é uma das limitações deste método uma vez que pode introduzir viés nos resultados.

Os parâmetros dos itens podem ser estimados recorrendo a métodos de máxima

verosimilhança ou a métodos de estimação bayesianos. Os métodos de máxima verosimilhança mais comuns são: máxima verosimilhança conjunta (por exemplo, Baker e Kim [11]), máxima verosimilhança marginal (por exemplo, Baker e Kim [11]) e máxima verosimilhança condicional (por exemplo, Baker e Kim [11]).

Com os parâmetros dos itens conhecidos os estimadores de máxima verosimilhança (EMV) do factor latente, pela propriedade da consistência, convergem em probabilidade para os valores verdadeiros dos parâmetros, e vice-versa, conhecido o factor latente, os EMV dos parâmetros dos itens convergem em probabilidade para os seus valores verdadeiros. O procedimento de máxima verosimilhança conjunta permite estimar simultaneamente o factor latente dos examinandos e os parâmetros de todos os itens, pela maximização conjunta da função de verosimilhança. Existem três desvantagens para a utilização deste procedimento. A primeira é que a curva característica do item é não linear, o que resulta em equações da verosimilhança não-lineares (Hambleton e Swaminathan [58]). A segunda é que para obter estimativas dos parâmetros dos itens mais exactas na aplicação do modelo logístico de 3 parâmetros é necessário que o número de examinandos seja grande (em geral, superior a 1000) (Lord e Novick [78]). A terceira é que o aumento do número de examinandos não garante uma melhoria na estimação, isto é, Wright [117] mostrou que quando os parâmetros dos itens são estimados conjuntamente com o factor latente, os EMV podem ser assimptoticamente enviesados.

O procedimento de máxima verosimilhança marginal (Baker e Kim [11]; Costa [24]) surgiu com o intuito de tentar solucionar problemas de falta de propriedades assimptóticas desejáveis nos EMV. Este procedimento baseia-se em considerar a existência de uma distribuição (latente) associada aos factores latentes dos examinandos da população em estudo. Deste modo, integra-se sobre a distribuição do factor latente e os parâmetros dos itens são estimados na distribuição marginal. Este procedimento permite solucionar o problema da dependência da estimação dos parâmetros dos itens sobre a estimação do factor latente. Para a obtenção das

estimativas de máxima verosimilhança, Bock e Aitkin [16] propuseram a aplicação do algoritmo *Expectation-Maximization* (EM), introduzido por Dempster, Laird e Rubin [35]. Uma das limitações que este procedimento apresenta é não estar definido para acerto total e erro total. A necessidade de estabelecer uma distribuição para o factor latente é outra das limitações deste procedimento.

O procedimento de máxima verosimilhança condicional pode ser usado para estimar os parâmetros dos itens e as estimativas obtidas são consistentes. Este procedimento permite estimar o factor latente, mas apenas para os modelos de Rasch. Outras desvantagens apontadas para o uso deste procedimento são: não está definido para alguns padrões de resposta, como itens com acerto total ou erro total, respostas omissas e itens politómicos, e, é frequente ocorrerem problemas numéricos em testes longos.

Para o procedimento de máxima verosimilhança são ainda apontadas as seguintes limitações: a possibilidade de que as estimativas dos parâmetros dos itens não pertençam ao intervalo esperado, tal como valores de $a$ negativos, ou valores de $c$ fora do intervalo [0,1]; em algumas situações, podem-se obter estimativas para o factor latente de infinito positivo ou negativo, e, o algoritmo EM torna-se difícil de aplicar quando as configurações do teste são mais complexas, como em dados omissos, itens politómicos e modelos de resposta ao item mais complexos, como é o caso dos modelos multidimensionais (Patz e Junker [92]).

A metodologia bayesiana apresenta uma solução para as limitações do procedimento de estimação pela máxima verosimilhança. Esta metodologia consiste em estabelecer valores à *priori* para os parâmetros de interesse (nos MRI, os parâmetros dos itens e do(s) factor(es) latente(s)), construir uma nova distribuição *a posteriori* e estimar os parâmetros de interesse com base em alguma característica dessa distribuição. Assume-se que $\beta$ é um vector de parâmetros. A distribuição *a posteriori* de $\beta$ é definida a partir do teorema de Bayes por:

$$p(\beta|U) \propto p(U|\beta)p(\beta) \tag{2.3.1}$$

onde $p(\beta)$ é a distribuição *a priori* para o vector de parâmetros $\beta$ e $U$ é a matriz de respostas.

Uma dificuldade na inferência bayesiana é obter a constante de normalização da equação anterior, uma vez que envolve integração numérica e nem sempre é possível encontrar uma forma analítica fechada para a distribuição *a posteriori*. Para colmatar esta limitação, usam-se alguns métodos baseados em simulação para aproximar essa distribuição.

### *Markov Chain Monte Carlo* em modelos de resposta ao item

A necessidade de utilizar um procedimento de estimação que permita a sua extensão, à medida que aumenta a complexidade dos modelos, fez com que se conjugasse a estimação bayesiana com o uso de métodos de simulação *Markov Chain Monte Carlo* (Patz e Junker [92] e [93]). Conceptualmente, a criação de um algoritmo de MCMC para um modelo de resposta ao item é semelhante ao algoritmo EM, mas como MCMC não envolve quadratura numérica exacta (para o passo E) ou o pré-cálculo das derivadas (para o passo M), então, MCMC é mais fácil de implementar. O método MCMC é usado para gerar as amostras da distribuição conjunta *a posteriori* dos parâmetros dos modelos de resposta ao item.

O primeiro autor a usar MCMC em modelos de resposta ao item foi Albert [4], que obteve as estimativas dos parâmetros do modelo de 2 parâmetros de ogiva normal para dados aumentados usando amostragem *Gibbs*. Patz e Junker ([92] e [93]) utilizaram MCMC em modelos unidimensionais de resposta ao item, em modelos que consideram respostas omissas e em modelos para itens politómicos. Em 2001, Béguin e Glas [12] generalizaram o procedimento exposto por Albert [4] para estimar os parâmetros do modelo de 2 parâmetros da ogiva normal. Estes autores apresentaram um procedimento de estimação para os parâmetros do modelo de 3

parâmetros da ogiva normal e expuseram uma generalização desse procedimento ao modelo multidimensional para o factor latente. O procedimento permite analisar dados para múltiplas populações e *designs* incompletos (subconjunto do conjunto total de itens é administrado a cada examinando). Contudo, é assumido que a matriz de covariância subjacente para identificar os factores latentes é uma matriz identidade, o que não é realista, uma vez que os factores latentes aferidos num teste, em geral, são correlacionados. De la Torre e Patz [113] propuseram a estimação simultânea dos factores latentes de modelos multidimensionais usando MCMC. No entanto, neste trabalho, os autores assumem que os parâmetros dos itens são conhecidos, o que na prática não costuma acontecer. Em 2003, Bolt e Lall [21] investigaram a estimação dos parâmetros dos itens de modelos de resposta ao item compensatórios e não compensatórios usando o algoritmo *Metropolis-Hastings* em MCMC. A limitação apontada para este trabalho é de ser apenas considerado o modelo bidimensional. Em 2005, Jiang [66] propôs usar *Metropolis Hastings* com amostragem *Gibbs* para efectuar a estimação simultânea dos parâmetros dos itens e do factor latente para o caso do modelo multidimensional logístico de 3 parâmetros. Este autor considerou os casos de 3 e 5 factores, utilizando apenas dados simulados, e modelou uma estrutura para a matriz de covariância que tornou a obtenção das estimativas dos parâmetros complexa computacionalmente e demorada.

Entretanto, autores (Baker [9]; Baker e Kim [11]; Kim [69]; Wollack *et al.* [116]) usaram em modelos de resposta ao item o método de simulação MCMC.

Neste trabalho, propomos utilizar MCMC conjugando *Metroplis-Hastings* com amostragem *Gibbs* para a obtenção das estimativas dos parâmetros dos itens e dos factores latentes, separadamente do modelo compensatório logístico multidimensional de 2 parâmetros. Este procedimento é testado usando dados simulados, dados reais e também considerando uma perspectiva de comparar os resultados obtidos com o software comercial mais utilizado para esta classe de modelos, Testfact 2.13 (Wilson, Wood e Gibbons [115]). O procedimento utilizado apresenta-se seguida-

mente.

### Procedimento de estimação proposto

No método de simulação MCMC geram-se amostras da distribuição de interesse a partir de distribuições que constituam uma cadeia de *Markov*. Tais distribuições são as chamadas distribuições de transição da cadeia que devem ser adequadamente escolhidas de forma a que a cadeia convirja em distribuição para a distribuição de interesse que, neste caso, corresponde à distribuição conjunta *a posteriori* dos parâmetros do modelo. Após atingir a convergência, as amostras são geradas a partir dessa distribuição estacionária. O objectivo é gerar uma amostra com uma dimensão suficientemente grande a partir desta distribuição estacionária que aproxime bem a distribuição conjunta *a posteriori*. Para isso, estipula-se um *burn-in* (número de iterações que se julgue serem necessárias para a convergência da cadeia).

Existem alguns métodos para a construção das cadeias, entre eles a amostragem *Gibbs* (Geman e Geman [52]) e que são casos especiais de uma estrutura geral do algoritmo *Metropolis-Hastings* (MH) (Metropolis *et al.* [86] e Hastings ([62]).

O método de amostragem *Gibbs* foi introduzido por Geman e Geman [52] e com grande contribuição de Gelfand e Smith [51]. O método consiste em considerar as distribuições condicionais completas dos parâmetros como as distribuições de transição da cadeia de *Markov*. Seja $\pi(\theta, X)$ a distribuição conjunta *a posteriori*, com $\theta = (\theta_1, \theta_2, ..., \theta_k)'$, onde cada componente de $\theta$ pode ser um escalar, um vector ou uma matriz. Dessa forma, as distribuições condicionais completas são $\pi_l(\theta_l) = \pi_l(\theta_l|\theta_{-l})$, $l = 1, 2, ..., k$ onde $\theta_{-l}$ é o vector $\theta$ sem a $l$-ésima componente. As amostras são geradas do seguinte modo:

1) Inicializa-se o contador das iterações da cadeia $s = 1$ e escolhem-se os valores iniciais para $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_k^{(0)})'$;

2) Obtém-se um novo valor $\theta^{(s)} = (\theta_1^{(s)}, \theta_2^{(s)}, ..., \theta_k^{(s)})'$ de $\theta^{(s-1)}$ através de gerações sucessivas dos valores:

$$\theta_1^{(s)} \sim \pi(\theta_1|\theta_2^{(s-1)}, ..., \theta_k^{(s-1)}),$$

$$\theta_2^{(s)} \sim \pi(\theta_2|\theta_1^{(s)}, \theta_3^{(s-1)}, ..., \theta_k^{(s-1)}),$$

$$...$$

$$\theta_k^{(s)} \sim \pi(\theta_k|\theta_1^{(s)}, ..., \theta_{k-1}^{(s)});$$

3) Muda-se o contador de $s$ para $s+1$ e retorna-se ao passo 2 até que seja gerado o tamanho de amostra desejado (considerando um *burn-in*).

É de notar que uma condição necessária da amostragem *Gibbs* é que as distribuições condicionais completas sejam conhecidas e que se saiba gerar valores dessas, porém, isso nem sempre é possível. Uma solução para este problema é o algoritmo MH. Neste caso, suponha-se que se deseja gerar uma amostra de uma distribuição de interesse através de cadeias de *Markov*, que, no caso, é a distribuição *a posteriori*, e defina-se $p(\theta, \phi)$ como a distribuição de transição que leva à convergência da cadeia para a distribuição de interesse. O algoritmo MH consiste em gerar os valores de uma distribuição de transição arbitrária com densidade $q(\theta, \phi)$, baseado numa probabilidade $\alpha(\theta, \phi)$ tal que:

$$p(\theta, \phi) = q(\theta, \phi)\alpha(\theta, \phi), \text{ se } \theta \neq \phi$$

Assim, $q(\theta, \phi)$ define uma densidade para $p(\theta, .)$ para todo o valor possível do parâmetro diferente de $\theta$. Consequentemente, existe uma probabilidade da cadeia permanecer em $\theta$ dada por:

$$p(\theta, \phi) = 1 - \int q(\theta, \phi)\alpha(\theta, \phi)d\phi$$

Baseado em Hastings [62], a expressão considerada para a probabilidade de aceitação é

$$\alpha(\theta, \phi) = min\left\{1, \frac{\pi(\phi)q(\phi,\theta)}{\pi(\theta)q(\theta,\phi)}\right\}$$

Nesse sentido, o algoritmo MH é o seguinte:

1) Inicializa-se o contador de iterações da cadeia $s = 1$ e escolhe-se um valor inicial arbitrário $\theta^{(0)}$;

2) Move-se a cadeia para um novo valor $\phi$ gerado da densidade $q(\theta^{(s-1)}, .)$ ;

3) Avalia-se a probabilidade de aceitação do movimento de transição $\alpha(\theta^{(s-1)}, \phi)$ dada pela expressão anterior. Se o movimento de transição é aceite, $\theta^{(s)} = \phi$; caso contrário, $\theta^{(s)} = \theta^{(s-1)}$ e a cadeia não se move.

4) Muda-se o contador de $s$ para $s + 1$ e retorna-se ao passo 2 até que seja gerado o tamanho da amostra desejado (considerando um *burn-in*).

Em modelos de resposta ao item, em geral, são conhecidas apenas algumas distribuições condicionais completas. Assim, neste trabalho, é adoptado o método conhecido como amostragem *Gibbs* com passos de *Metropolis* proposto por Muller [87]. Como as componentes não podem ser geradas das amostras directamente a partir das respectivas condicionais completas $\pi_i$, são geradas as amostras de $\pi_i$ através de uma sub-cadeia de *Metropolis-Hastings* dentro do ciclo amostral de *Gibbs*. Essas componentes permitem gerar uma amostra da proposta $q_i$ e serem aceites com probabilidade $\alpha(\theta, \phi)$. Quanto mais parecidas forem a proposta $q_i$ e a condicional completa $\pi_i$, mais próxima de 1 será a probabilidade de aceitação.

Seguidamente, propõe-se o procedimento para se fazer inferência sobre o modelo logístico multidimensional de 2 parâmetros, utilizando esta abordagem. Para isso, apresentam-se as seguintes especificações: função de verosimilhança, distribuições *a priori*, distribuição conjunta *a posteriori*, distribuições condicionais completas e o algoritmo de aplicação de MCMC.

### Função de verosimilhança

A função de verosimilhança (Baker e Kim [11]) é dada por:

$$L(U|\Theta, A, d) = \prod_{j=1}^{J} \prod_{i=1}^{I} p_i(\theta_j)^{U_{ij}} (1 - p_i(\theta_j))^{(1-U_{ij})}$$

onde,

$i$ representa o item, com $i = 1, 2, ..., I$;

$j$ representa o examinando, com $j = 1, 2, ..., J$;

$\theta_j$ é um vector com $k$ componentes, que representa os $k$ factores latentes referentes ao examinando $j$, isto é, $\theta_j = (\theta_{j1}, \theta_{j2}, ..., \theta_{jk})$

$p_i(\theta_j)$ representa a probabilidade do examinando $j$ responder correctamente ao item $i$, dada pelo modelo multidimensional logístico de 2 parâmetros, isto é, $p_i(\theta_j) = P(U_{ij} = 1|\theta_j, a_i, d_i)$;

$a_i$ é um vector com $k$ componentes, que representa o parâmetro de discriminação do item $i$ específico de cada dimensão dos $k$ factores latentes, isto é, $a_i = (a_{i1}, a_{i2}, ..., a_{ik})$;

$d_i$ é o parâmetro de dificuldade do item $i$;

$U_{ij}$ representa a resposta dicotómica do examinando $j$ ao item $i$, em que $U_{ij} = 1$, se for resposta correcta e $U_{ij} = 0$, se for resposta incorrecta;

$U$ é a matriz de respostas dos $J$ examinandos aos $I$ itens;

$\Theta$ é uma matriz $J \times k$ que representa todos os factores latentes aferidos no instrumento, isto é, $\Theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \dots & \theta_{1k} \\ \theta_{21} & \theta_{22} & \dots & \theta_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{J1} & \theta_{J2} & \dots & \theta_{Jk} \end{bmatrix}$

$A$ é uma matriz $I \times k$ que representa todos os parâmetros $a$ para os $I$ itens, isto é, $A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ a_{I1} & a_{I2} & \dots & a_{Ik} \end{bmatrix}$

$d$ é o vector do parâmetro de dificuldade para os $I$ itens, isto é, $d = (d_1, d_2, ..., d_I)$;

### Distribuições *a priori*

Assume-se que os parâmetros são independentes à *priori*, o que significa que $a$

*priori* conjunta (Baker e Kim [11]) é dada por:

$$\pi(\Theta, A, d) = \pi_\Theta(\Theta)\pi_A(A)\pi_d(d)$$

ou

$$\pi(\Theta, A, d) = \sum_{j=1}^{J} \pi_\Theta(\Theta) \sum_{i=1}^{I} \pi_A(A)\pi_d(d)$$

Para cada componente, especificam-se as seguintes distribuições *a priori*:

- $a_i \sim NM(\mu_a, \Omega_a)$ onde $NM(.,.)$ representa a distribuição normal multivariada com vector média $\mu_a$ e matriz de covariâncias $\Omega_a$ ($\mu_a$ é um vector com $k$ zeros e $\Omega_a$ é a matriz identidade $I_k$;

- $d_i \sim N(\mu_d, c_d^2)$, onde $N(.,.)$ representa a distribuição normal com $\mu_d = 0$ e $c_d^2 = 2$ (escolhido de forma a que os valores obtidos pertençam ao intervalo em que se espera encontrar os valores reais da dificuldade dos itens);

- $\theta_j \sim NM(\mu_\theta, \Omega_\theta)$ onde $\mu_\theta$ é o vector nulo com $k$ componentes e $\Omega_\theta$ é a matriz identidade $I_k$;

### Distribuição conjunta *a posteriori*

A distribuição conjunta *a posteriori* é utilizada para obter as estimativas de todos os parâmetros do modelo e foi obtida pela aplicação do teorema de Bayes. Assume-se que as distribuições *a priori* para os itens e os factores latentes são independentes. A distribuição conjunta *a posteriori* (Baker e Kim [11]) é dada por:

$$p(A, d, \Theta|U) \propto L(U|\Theta, A, d)\pi(\Theta, A, d)$$

Em geral, é difícil obter a constante de normalização para a expressão anterior, o que significa que não é possível encontrar uma forma analítica fechada para a distribuição *a posteriori*.

### Distribuições condicionais completas

A distribuição condicional completa (Baker e Kim [11]) para os parâmetros referentes aos factores latentes dos examinandos, pela aplicação do teorema de Bayes, é:

$$P_\theta(\theta_j|\Theta_{-j}, A, d, U) = \prod_{i=1}^{I} p_i(\theta_j)^{U_{ij}}(1 - p_i(\theta_j))^{(1-U_{ij})}\pi_\theta(\theta_j)$$

onde, $\Theta_{-j} = (\theta_1, \theta_2, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_J)'$.

De modo análogo se obtêm as distribuições condicionais completas para os parâmetros dos itens. Isto é,

$$P_a(a_i|A_{-i}, \Theta, d, U) = \prod_{j=1}^{J} p_i(\theta_j)^{U_{ij}}(1 - p_i(\theta_j))^{(1-U_{ij})}\pi_a(a_i)$$

e

$$P_d(d_i|d_{-i}, \Theta, A, U) = \prod_{j=1}^{J} p_i(\theta_j)^{U_{ij}}(1 - p_i(\theta_j))^{(1-U_{ij})}\pi_d(d_i)$$

onde,

$A_{-i}$ é uma matriz $(I-1) \times k$, isto é, $A_{-i} = (a_1, a_2, ..., a_{i-1}, a_{i+1}, ..., a_I)'$;

$d_{-i}$ é um vector com $(I-1)$ componentes, isto é, $d_{-i} = (d_1, d_2, ..., d_{i-1}, d_{i+1}, ..., d_I)$.

### Valores iniciais

Os valores iniciais adoptados para:

i) os parâmetros de discriminação dos itens - $a_i^0 = 1$;

ii) o parâmetro de dificuldade dos itens - $d_i^0 = 0$;

iii) para todos os factores latentes - $\theta_j^0 \sim N(0, 1)$.

### Algoritmo

O algoritmo para a estimação dos parâmetros do modelo multidimensional logístico de 2 parâmetros foi implementado em Matlab e apresenta-se seguidamente.

**Passo 1.** Inicializa-se o contador de iterações da cadeia $s = 1$;

**Passo 2.** Gera-se, $\forall j = 1, ..., J$, os factores latentes $\theta_j^s$.

Para obter $\theta_j^s$, a partir de $p(\theta|A^{s-1}, d^{s-1}, U)$:

**a)** Gera-se $\theta_j^* \sim NM(\mu_{\theta_j}|\mu_{\theta_j^{s-1}}, \Omega_{\theta_j})$ independentemente para cada $j = 1, ..., J$;

**b)** Calculam-se os vectores de $J$ probabilidades de aceitação, $\alpha(\theta_j^{s-1}, \theta_j^*) = min\{1, r_\theta\}$ onde $r_\theta = \frac{L(U_j|\theta_j^*, A^{s-1}, d^{s-1})\pi(\theta_j^*)}{L(U_j|\theta_j^{s-1}, A^{s-1}, d^{s-1})\pi(\theta_j^{s-1})}$ para $j = 1, ..., J$;

**c)** Aceita-se $\theta_j^s = \theta_j^*$ com probabilidade $\alpha_j$, caso contrário tem-se $\theta_j^s = \theta_j^{s-1}$.

**Passo 3.** Gera-se $\forall i = 1, ..., I$, os parâmetros $d_i^s$.

Para obter $d^s$ a partir de $p(d|\theta^{s-1}, A_{s-1}, U)$:

**a)** Gera-se $d_i^* \sim N(d_i|d_i^{s-1}, c_{d_i}^2)$ independentemente para cada $i = 1, ..., I$ (Em particular foi considerado $c_{d_i}^2 = 0,01$);

**b)** Calcula-se o vector de $I$ probabilidades de aceitação $\alpha(d_i^{s-1}, d_i^*) = min\{1, r_d\}$, onde $r_d = \frac{L(U_i|d_i^*, \theta^{s-1}, A^{s-1})\pi(d_i^*)}{L(U_i|d_i^{s-1}, \theta^{s-1}, A^{s-1})\pi(d_i^{s-1})}$ para $i = 1, ..., I$;

**c)** Aceita-se $d_i^s = d_i^*$ cada com probabilidade $\alpha_i$, caso contrário faz-se $d_i^s = d_i^{s-1}$.

**Passo 4.** Gera-se $\forall i = 1, ..., I$, os parâmetros $a_i^s$.

Para obter $a^s$ a partir de $p(A|\theta^{s-1}, d^{s-1}, U)$:

**a)** Gera-se $a_i^* \sim NM(\mu_{a_i}|\mu_{a_i^{s-1}}, \Omega_{a_i})$ independentemente para cada $i = 1, ..., I$;

**b)** Calculam-se os vectores de $I$ probabilidades de aceitação, $\alpha(a_i^{s-1}, a_i^*) = min\{1, r_a\}$, onde $r_a = \frac{L(U_i|a_i^*, \theta^{s-1}, d^{s-1})\pi(a_i^*)}{L(U_i|a_i^{s-1}, \theta^{s-1}, d^{s-1})\pi(a_i^{s-1})}$ para $i = 1, ..., I$;

**c)** Aceita-se $a_i^s = a_i^*$ com probabilidade $\alpha_i$, caso contrário tem-se $a_i^s = a_i^{s-1}$.

**Passo 5.** Muda-se o contador de $s$ para $s+1$ e retorna-se ao passo 2 até que seja gerado o tamanho da amostra desejado (considerando um *burn-in*).

No anexo 1 encontra-se o algoritmo implementado em Matlab, usado para o obtenção das estimativas dos parâmetros do MML2. O fluxograma e o ciclo que ilustram o algoritmo descrito apresenta-se no anexo 2.

## 2.4 Considerações gerais

Os modelos de resposta ao item multidimensionais podem ser usados para verificar o pressuposto de unidimensionalidade dos MRI unidimensionais. Seguidamente vamos apresentar os métodos existentes para a análise da dimensionalidade de instrumentos.

**Análise da dimensionalidade**

A questão da dimensionalidade de um teste/instrumento consiste em verificar quantos factores estão a ser medidos. A realização deste tipo de análise permite, por exemplo, a verificação do pressuposto de que existe um factor latente dominante responsável pelo desempenho num conjunto de itens de um teste (unidimensionalidade) e deve ser verificado em cada teste para se poder utilizar qualquer um dos modelos de resposta ao item unidimensionais.

Existem métodos clássicos usados para analisar a dimensionalidade, nomeadamente, a correlação bisserial (D'Hainaut [36]) e a correlação tetracórica (D'Hainaut [36]). O desenvolvimento dos modelos de resposta ao item possibilitou o aparecimento de novos métodos para a análise da dimensionalidade. Segundo Soares [106], existem dois tipos de métodos para a análise da dimensionalidade associada a um conjunto de variáveis dicotómicas: métodos da informação restrita e os métodos de informação plena.

O método de informação restrita (Soares [106]) consiste na inspecção dos valores próprios da matriz de correlação tetracórica (D'Hainaut [36]) quanto aos demais. Um método destinado a obter uma aproximação da correlação tetracórica é apresentado em Divgi [37], e está disponível no software Testfact 2.13 (Wilson, Wood e Gibbons [115]).

Os métodos de análise factorial da informação plena (Bock e Aitkin [16]; Bock, Gibbons e Muraki [17]; Muraki e Engelhard [91]), foram propostos a partir de uma adaptação do modelo tradicional de análise factorial (Johnson e Wicherin [67] e Johnson e Wicherin [68]) que considera a estrutura de dimensões associadas a variáveis contínuas (Thurstone [111]).

*Modelo de análise factorial para variáveis dicotómicas*

Segundo Soares [106], ambos os métodos de análise da dimensionalidade surgiram a partir do modelo de análise factorial, considerando a estrutura de dimensões associadas a variáveis contínuas. Assim, a definição de uma variável artificial é a chave para a construção do método. Nesse sentido, definindo uma variável $X_i$, tal que $\sigma_X = 1$, e $E(X_i) = 0$, e a relação dessa variável com a variável dicotómica $U_i$ que representa a resposta atribuída ao item $i$ (assumindo os valores 0 ou 1) é tal que:

$$\text{Se } X_i \geq \gamma_i, \text{ então } U_i = 1$$

e,

$$\text{Se } X_i < \gamma_i, \text{ então } U_i = 0.$$

O modelo de análise factorial é então definido a partir da variável $X_i$ da seguinte forma:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1d} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2d} \\ \vdots & & & \\ \lambda_{n1} & \lambda_{n2} & \dots & \lambda_{nd} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

ou

$$X = \Lambda\theta + \epsilon \tag{2.4.2}$$

Os valores $\lambda_{ij}$ são as cargas associadas ao factor $\theta_j$ e à variável $X_i$, medida do grau de associação entre o factor e a variável. Considerando os pressupostos para o modelo de factores ortogonal, a correlação de $X$ é dada por:

$$\Sigma = \Lambda\Theta\Lambda' + \Psi$$

onde $\Theta$ é a matriz de covariância de $\theta$ e $\Psi$ é tal que $\epsilon \sim N(0, \Psi)$, com $\Psi$ diagonal. Em particular, se o modelo é unidimensional, então as linhas de $\Lambda\Theta\Lambda'$ serão todas linearmente dependentes entre si e, portanto, os seus valores próprios serão todos iguais a zero, excepto um deles. Na prática, a unidimensionalidade deve ser entendida como a predominância de uma única dimensão sobre as demais. Assim, o primeiro método para se testar a dimensionalidade, que emerge naturalmente nesse contexto, é o da inspecção dos valores próprios da matriz de correlações tetracóricas, considerando-se a dimensão do modelo o número de valores próprios superiores a um determinado valor (normalmente, um). Mas esse critério é altamente subjectivo. De facto, na prática aceita-se como a dimensão associada às variáveis, um certo número de valores próprios, cujos valores sejam razoavelmente maiores que os demais, embora esse critério seja também subjectivo.

*Método de análise factorial de informação plena*

Com o intuito de evitar a subjectividade inerente ao uso do método da informação restrita para a detecção da dimensionalidade, Bock e Aitkin [16], Bock, Gibbons e Muraki [17], propuseram o método da análise factorial de informação plena. Considere-se, novamente, o modelo de análise factorial apresentado anteriormente. Dessa forma, então, $P(U_i = 1) = P(X_i \geq \gamma_i) = P(\sum_{j=1}^{d} \lambda_{ij}\theta_j + \epsilon_i \geq \gamma_i) = P(\epsilon_i \geq \gamma_i - \sum_{j=1}^{d} \lambda_{ij}\theta_j)$. Relembrando que por hipótese $\epsilon \sim N(0, \Psi)$, com $\Psi$ diagonal, tem-se então que:

$$P(U_i = 1) = \int_{\frac{\gamma_i - \sum_{j=1}^{d} \lambda_{ij}\theta_j}{\sigma_{e_i}}}^{\infty} \frac{e^{\frac{-z^2}{2}}}{\sqrt{2\pi}} dz \tag{2.4.3}$$

onde $\sigma_{e_i}^2$ é a variância de $\epsilon_i$. Da estrutura do modelo pode-se verificar que $\sigma_{e_i} = \sqrt{1 - \sum_{j=1}^{d} \lambda_{ij}^2}$, e reparametrizando a fórmula anterior do seguinte modo:

$$d_i = -\frac{\gamma_i}{\sigma_{e_i}}, a_i = \frac{\lambda_{ij}}{\sigma_{e_i}}, \tag{2.4.4}$$

tem-se um modelo multidimensional que utiliza a curva de ogiva normal com $d_i$ como sendo a dificuldade geral do item e os valores de $a_i$ como sendo os parâmetros de discriminação específicos de cada dimensão. Considerando a probabilidade de acerto ao acaso no caso multidimensional, a equação do modelo corresponde à equação 2.2.5.

O método para estimação dos parâmetros desse modelo pode ser, *mutato mutandis*, o mesmo método de máxima verossimilhança marginal (Bock e Aitkin [16]) utilizado nos modelos mais comuns. Nota-se que as equações 2.4.4 fornecem uma forma directa para se obter as cargas do modelo de análise factorial, para isso, basta que se invertam as relações. Para a estimação dos parâmetros do modelo, empregando-se o método mencionado, utiliza-se o software Testfact 2.13 (Wilson, Wood e Gibbons [115]). O parâmetro de probabilidade de acerto ao acaso, é calculado recorrendo ao software Bilog-MG 3.0 (Zimowski *et al.* [118]) e, posteriormente, é inserido como constante no software Testfact 2.13 (Wilson, Wood e Gibbons [115]).

Neste capítulo, apresentámos a extensão dos modelos de resposta ao item unidimensionais aos modelos multidimensionais. Seguidamente, expusémos os principais procedimentos de estimação para estimar os parâmetros dos modelos multidimensionais. Propusémos um procedimento de estimação bayesiano, usando MCMC, para obter as estimativas dos parâmetros do modelo multidimensional logístico de 2 parâmetros. Apresentámos, ainda, os dois métodos que existem para a análise da dimensionalidade de testes: métodos de informação restrita e de informação plena. Os modelos multidimensionais permitem estabelecer as relações entre os examinandos

e as características dos itens do instrumento, considerando que se aferem múltiplos factores no instrumento. Esta classe de modelos pode, ainda, ser usada para verificar o pressuposto da unidimensionalidade dos modelos unidimensionais. No capítulo 4, apresentam-se várias aplicações dos modelos multidimensionais de resposta ao item, tanto a dados reais como a dados simulados.

# Capítulo 3

# Aplicações - Modelos de Resposta ao Item Unidimensionais

Neste capítulo, apresentamos aplicações de modelos de resposta ao item unidimensionais dicotómicos, politómicos e modelos para grupos múltiplos. Adicionalmente, utilizamos procedimentos de equalização e *linking* para estabelecer a comparação de classificações obtidas em diferentes instrumentos.

Começamos por estudar as propriedades psicométricas de um instrumento/teste de escolha múltipla, utilizado para aferir competências em Estatística, no âmbito do curso de Mestrado Integrado em Engenharia e Gestão Industrial da Universidade do Minho. Para realizar a análise dos dados, recorremos às abordagens baseadas na TCT e nos MRI. Em termos de MRI, usamos o modelo dicotómico unidimensional logístico de 2 parâmetros, apresentado na subsecção 1.4.2 (equação 1.4.2). Comparamos os resultados em dois anos lectivos, considerando cada conjunto de alunos como amostras independentes da mesma população.

Na secção 3.2 exploramos a utilização de modelos de resposta ao item politómicos unidimensionais. Modelamos os dados da prova de aferição de Matemática do 4º ano de escolaridade, do 1º Ciclo do Ensino Básico, aplicada no ano lectivo 2006/2007, usando o MRI de Crédito Parcial Generalizado. Esta aplicação tem como objec-

tivos: estudar as propriedades estatísticas dos instrumentos de medida dos resultados escolares do 4º ano de escolaridade no ano lectivo 2006/2007, a identificação de "perguntas-problema" e a quantificação do impacto das "perguntas-problema" nos resultados escolares. Para validar a metodologia de análise avaliamos a adequação do modelo teórico aos dados. Com vista à análise dos itens que compõem a prova, apresentamos as estimativas do parâmetro de discriminação, do parâmetro de dificuldade e dos parâmetros de intersecção das categorias adjacentes. Para averiguar a capacidade informativa do teste relativamente ao factor latente (competências em Matemática no 4º ano), recorremos à função de informação do teste e quantificamos o erro da medida.

Usamos, de seguida, o modelo de resposta ao item unidimensional para grupos múltiplos com vista a obter numa escala única as estimativas dos parâmetros dos itens de testes que aferem aprendizagens a Matemática. Esta aplicação faz parte do procedimento usado para analisar a dimensionalidade de testes que aferem aprendizagens a Matemática. Os dados referem-se aos testes dos 1º, 2º e 3º anos de escolaridade do Ensino Básico e foram aplicados no âmbito do projecto Eficácia Escolar no Ensino da Matemática - 3EM. Recorremos ao procedimento de estimação de máxima verosimilhança para a obtenção das estimativas dos parâmetros do modelo para grupos múltiplos.

Na secção 3.4, usamos o procedimento estatístico de equalização com o objectivo de ajustar uma escala vertical única padronizada do desempenho na disciplina de Matemática para o Ensino Básico. A equalização permite proceder à comparação das competências desenvolvidas em diferentes populações de alunos ao longo do tempo, bem como comparar as competências desenvolvidas ao longo da sua trajectória escolar (equalização vertical). Os testes, designados 3EMat, foram recolhidos no âmbito do projecto 3EM. As escalas de desempenho, para aferir as competências desenvolvidas em Matemática, nos diferentes níveis do Ensino Básico, são feitas com base na aplicação do modelo de resposta ao item logístico de 2 parâmetros. O

procedimento de estimação de Máxima Verosimilhança Marginal (MVM) é utilizado para estimar o factor latente, desempenho em Matemática.

Na última secção, com o propósito de aplicar o procedimento estatístico *linking*, efectuamos a ligação entre as escalas construídas a partir o teste 3EMat e a prova de aferição, considerando a disciplina de Matemática do 6º ano (PAM6) de escolaridade aplicada no ano lectivo 2006/2007 (Anexo 5). Para isso, utilizamos o método linear e a estimação conjunta, assumindo que cada um dos instrumentos (PAM6 e 3Emat) são subtestes aplicados à mesma amostra. Realizamos a comparação dos resultados das distribuições marginais das classificações e o estudo da distribuição conjunta, no que se refere ao *matching* de casos válidos, e comparamos as escalas PAM6 e 3EMat em termos dos resultados obtidos pelos alunos em ambos os instrumentos.

## 3.1 Dicotómicos

### Dados e Resultados

Os dados foram recolhidos no âmbito do projecto da Universidade do Minho e estão descritos no Anexo 3 (secção 3.1). Nesta aplicação, modelam-se os dados do 1º teste, que afere competências em Estatística Descritiva e que é composto por 19 itens.

A interpretação de cada item do teste foi feita, inicialmente, a partir das estatísticas da TCT (Costa, Oliveira e Ferrão [30]): índice de dificuldade, índice de discriminação e correlação ponto-bisserial.

As tabelas 3.1 e 3.2 apresentam a classificação dos itens do teste, segundo os índices de dificuldade e de discriminação.

Os resultados mostram que, em geral, o teste é composto, essencialmente, por itens discriminativos e muito discriminativos e que existem itens de todos os níveis de dificuldade.

Tabela 3.1: Classificação dos itens segundo o índice de dificuldade

| Índice de dificuldade | TCT |
|---|---|
| Muito Difícil | 1 |
| Difícil | 5 |
| Médio | 4 |
| Fácil | 6 |
| Muito Fácil | 3 |

Tabela 3.2: Classificação dos itens segundo o índice de discriminação

| Índice de discriminação | TCT |
|---|---|
| Pouco discriminativo | 4 |
| Discriminativo | 2 |
| Muito discriminativo | 13 |

Os índices de dificuldade e de discriminação e o coeficiente de correlação ponto-bisserial obtidos para os 19 itens que compõem o teste são apresentados na tabela 3.3.

A análise da tabela permite verificar que os itens mais difíceis são 1, 13, 16 e 17 (valores mais baixos do índice de dificuldade) e os mais fáceis são o 4, 12 e 15, que correspondem aos valores mais elevados do índice de dificuldade. Os itens mais discriminativos são 4, 7, 8, 15, 18 e 19 e os itens menos discriminativos são 1, 6, 10 e 16. No que se refere ao item 1, o resultado obtido (índice de discriminação aproximadamente zero) chama a atenção para a situação em que a proporção de acerto no grupo de alto desempenho é muito próxima à do grupo de baixo desempenho. O valor do coeficiente de correlação ponto-bisserial associado a cada um dos itens permite constatar que os itens que mais se correlacionam com a classificação total são 4 e 15 e os que menos se correlacionam são 1, 10 e 16. A correlação negativa que ocorre com o item 1 corrobora a análise anterior referente ao índice de discriminação. Nestes termos, o item 1 deve ser revisto ou até retirado do teste.

A figura 3.1 apresenta o gráfico de dispersão dos índices em análise onde se pode observar que os itens 1 e 16 são itens difíceis, mesmo para o grupo de alto

Tabela 3.3: Índices de discriminação e de dificuldade e correlação ponto-bisserial para os itens que compõem o teste

| Item | Índice de dificuldade | Índice de discriminação | Correlação Ponto-bisserial |
|---|---|---|---|
| 1 | 0,120 | 0,024 | -0,005 |
| 2 | 0,354 | 0,333 | 0,305 |
| 3 | 0,677 | 0,524 | 0,438 |
| 4 | 0,753 | 0,571 | 0,521 |
| 5 | 0,348 | 0,548 | 0,447 |
| 6 | 0,601 | 0,262 | 0,266 |
| 7 | 0,513 | 0,571 | 0,375 |
| 8 | 0,475 | 0,667 | 0,487 |
| 9 | - | - | - |
| 10 | 0,620 | 0,167 | 0,178 |
| 11 | 0,500 | 0,476 | 0,389 |
| 12 | 0,785 | 0,333 | 0,381 |
| 13 | 0,272 | 0,429 | 0,422 |
| 14 | 0,576 | 0,524 | 0,431 |
| 15 | 0,759 | 0,571 | 0,532 |
| 16 | 0,253 | 0,095 | 0,084 |
| 17 | 0,304 | 0,548 | 0,492 |
| 18 | 0,557 | 0,571 | 0,467 |
| 19 | 0,525 | 0,643 | 0,490 |
| 20 | 0,582 | 0,500 | 0,360 |

desempenho e o item 8 aparenta ser um item fortemente discriminativo e de dificuldade média. Evidenciam-se também os itens 4 e 15, que são itens discriminativos e simultaneamente fáceis. O que acontece é que nos itens 4 e 15 no grupo de alto desempenho a proporção de acerto ao item é 1 e 0,98, no grupo de baixo desempenho é 0,43 e 0,4 e entre os restantes é 0,8 e 0,84, respectivamente.

A escala produzida para os 19 itens apresenta consistência interna medida pelo coeficiente de correlação KR20 (Capítulo 1 - secção 1.1) de 0,67 (SEM=2,295). Para avaliar o impacto dos itens 1, 10 e 16, calculou-se a classificação total do teste retirando estes três itens, o coeficiente de correlação KR20 para os 16 itens variou para 0,72 (SEM=2,089).

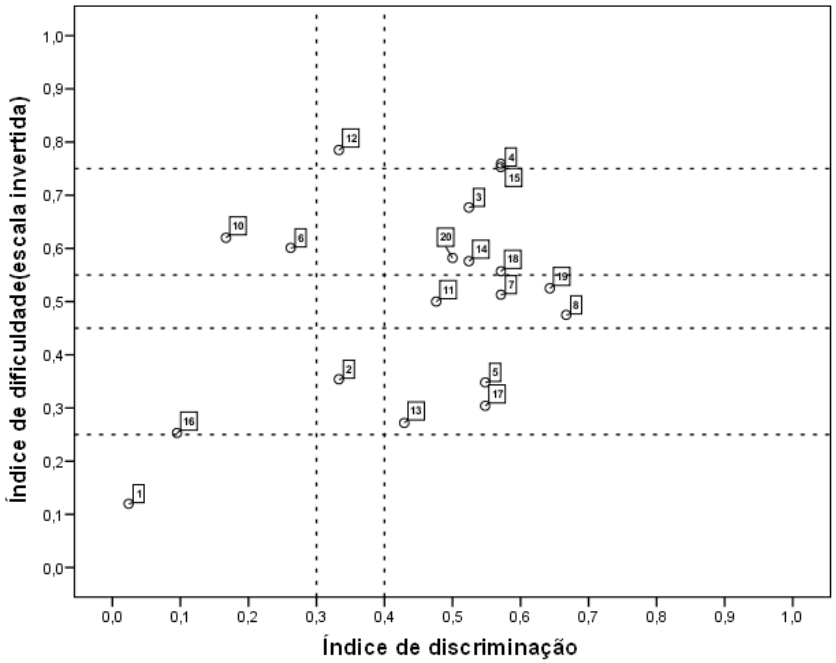



Figura 3.1: Diagrama de dispersão dos itens, índice de discriminação por índice de dificuldade

Para os outros dois testes analisados no âmbito do projecto, que aferem a aprendizagem dos conteúdos Probabilidades e Distribuições de Probabilidade (Costa, Oliveira e Ferrão [30]), o valor do coeficiente de consistência interna associado a cada um desses testes pertence ao intervalo [0,76; 0,82], tanto considerando todos os itens como excluindo os piores.

No que se refere à análise dos resultados pela aplicação de MRI (Ferrão, Costa e Oliveira [45]), os dados foram modelados pela aplicação do modelo logístico de 2 parâmetros (Birnbaum [14]; Hambleton, Swaminathan e Rogers [59]; Costa e Ferrão [25]; Costa, Oliveira e Ferrão [28]), apresentado no capítulo 1, subsecção 1.4.2. O procedimento de estimação de Máxima Verosimilhança Marginal (MVM) (Baker e Kim [11]) foi usado para obter as estimativas do factor latente e as estimativas dos parâmetros de discriminação e dificuldade dos itens. A estimação que recorre a este procedimento foi efectuada no programa computacional Bilog-MG (Zimowski *et al.*

[118]; Toit [112]), que é o software comercial utilizado e distribuído pelo Scientific Software International (SSI).

As tabelas 3.4 e 3.5 mostram a classificação dos itens do teste de acordo com as estimativas dos parâmetros de dificuldade ($\widehat{b}$) e de discriminação ($\widehat{a}$), respectivamente. Os itens 1 e 16 foram retirados pelo software Bilog-MG na fase 1, uma vez que apresentaram valores de correlação bisserial negativos, $-0,147$ e $-0,053$, respectivamente. Assim, daqui em diante serão considerados apenas os restantes 17 itens do teste.

Tabela 3.4: Classificação dos itens face à estimativa do parâmetro de dificuldade

| $\widehat{b}$ | MRI |
|---|---|
| Difícil | 2 |
| Médio | 10 |
| Fácil | 5 |

Tabela 3.5: Classificação dos itens face à estimativa do parâmetro de discriminação

| $\widehat{a}$ | MRI |
|---|---|
| Pouco discriminativo | 3 |
| Discriminativo | 7 |
| Muito discriminativo | 7 |

Os resultados vêm corroborar que o teste é constituído por itens essencialmente discriminativos e muito discriminativos, e a existência de itens de todos os níveis de dificuldade.

As estimativas dos parâmetros de dificuldade e de discriminação obtidas para todos os itens são apresentadas na tabela 3.6.

Relativamente às estimativas do parâmetro de dificuldade, constata-se que os itens mais difíceis são 2 e 13 e os itens mais fáceis são 3, 4, 6, 10 e 12. Os itens com estimativas mais elevadas do parâmetro de discriminação são 4 e 15, enquanto que os itens menos discriminativos são 6, 7 e 10. A representação gráfica da dispersão dos itens considerando as estimativas dos parâmetros dos itens (figura 3.2) ilustra

Tabela 3.6: Estimativas dos parâmetros de dificuldade e de discriminação dos itens que compõem o teste

| Item | $\widehat{b}$ | $\widehat{a}$ | Item | $\widehat{b}$ | $\widehat{a}$ |
|---|---|---|---|---|---|
| 1 | - | - | 11 | - | - |
| 2 | 0,765 | 0,524 | 12 | -1,11 | 0,883 |
| 3 | -0,804 | 0,572 | 13 | 1,321 | 0,534 |
| 4 | -0,823 | 1,268 | 14 | -0,245 | 0,794 |
| 5 | 0,727 | 0,682 | 15 | -0,688 | 1,468 |
| 6 | -1,119 | 0,258 | 16 | - | - |
| 7 | -0,345 | 0,196 | 17 | 0,723 | 0,846 |
| 8 | 0,083 | 0,708 | 18 | -0,154 | 0,838 |
| 9 | - | - | 19 | -0,217 | 0,503 |
| 10 | -1,078 | 0,24 | 20 | -0,395 | 0,486 |

que o teste é constituído, essencialmente, por itens discriminativos, de nível de dificuldade média e fácil. Os itens menos discriminativos (6, 7 e 10) são fáceis e os mais discriminativos, 4 e 15, são de dificuldade baixa e média, respectivamente.

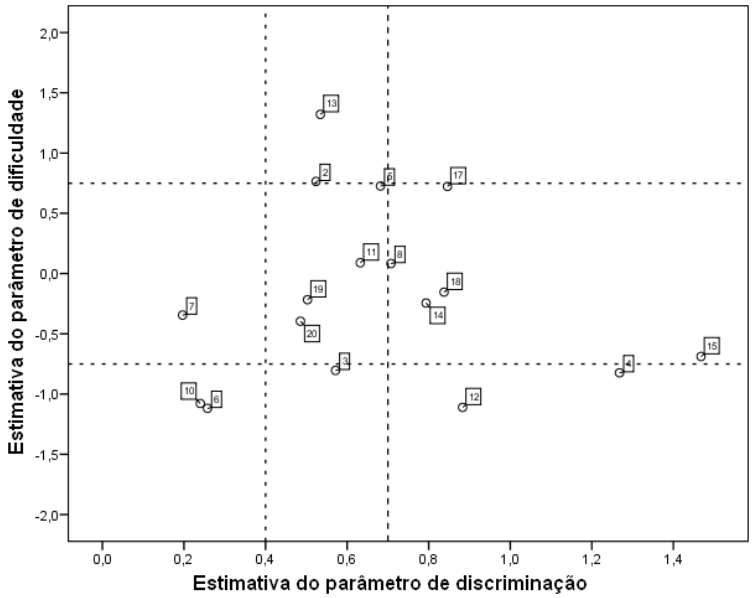



Figura 3.2: Diagrama de dispersão dos itens, estimativa do parâmetro de discriminação por estimativa do parâmetro de dificuldade

A função de informação do teste (figura 3.3) indica que os dados recolhidos através deste teste contêm grande poder informativo relativamente ao factor latente de alunos com nível de conhecimento compreendido entre -1 e 1. O erro padrão da medida indica elevada imprecisão dos resultados obtidos no que se refere à aferição da aprendizagem de alunos com nível de conhecimento nos extremos da escala.

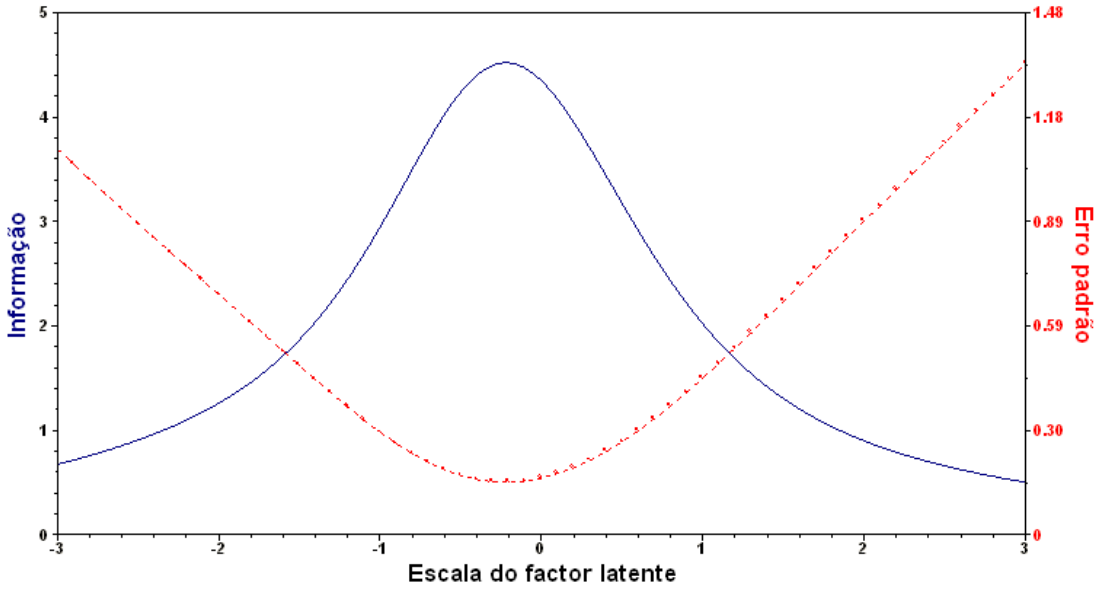



Figura 3.3: Função de informação do teste e erro padrão da medida

A análise dos itens que compõem o teste, considerando os resultados obtidos via TCT e os obtidos via MRI, permite verificar que os itens 1, 10 e 16 são os piores do teste e os itens 4 e 15 são os melhores.

As estatísticas descritivas da classificação dos alunos, nos anos lectivos 2006/2007 e 2007/2008, são apresentadas na tabela 3.7.

Constata-se que a distribuição das classificações resultantes da aplicação do teste em ambos os anos lectivos é simétrica. A amplitude interquartis da distribuição das classificações obtidas pela aplicação do teste no ano lectivo 2006/2007 é 4 e no ano lectivo 2007/2008 é 3. No ano lectivo 2007/2008 todas as medidas de localização apresentadas são superiores às do ano lectivo anterior.

Tabela 3.7: Estatísticas descritivas das classificações nos anos lectivos 2006/2007 e 2007/2008

| | Classificações 2006/2007 | Classificações 2007/2008 |
|---|---|---|
| N | 94 | 64 |
| Média | 8,085 | 11,766 |
| Mediana | 8 | 12 |
| Desvio padrão | 3,043 | 2,549 |
| Coeficiente de Assimetria | 0,240 | 0,17 |
| Erro padrão da Assimetria | 0,249 | 0,299 |
| Mínimo | 2 | 6 |
| Máximo | 16 | 18 |
| Percentil 25 | 6 | 10 |
| Percentil 75 | 10 | 13 |

Para verificar se existem diferenças das médias entre as classificações obtidas pela aplicação do teste aos alunos do ano lectivo 2006/2007 e do ano lectivo 2007/2008 foi aplicado um teste de hipóteses para amostras independentes. A estatística de teste tem associado um valor de prova (p=0,000) p<0,001, pelo que se conclui que existe diferença estatisticamente significativa entre as classificações obtidas pela aplicação do teste aos alunos dos anos lectivos 2006/2007 e 2007/2008.

### Discussão

A análise dos dados baseada nas abordagens TCT e MRI indicou que o teste é constituído, essencialmente, por itens discriminativos de todos os níveis de dificuldade e que os itens 1, 10 e 16 são os piores do teste e os itens 4 e 15 são os melhores. A escala produzida apresentou consistência interna adequada. Este teste contribuiu com maior informação para alunos de nível de factor latente médio. Para comparar as médias das classificações dos alunos, em ambos os anos lectivos, aplicámos o teste de hipóteses para amostras independentes e constatámos que existia uma melhoria nas médias das classificações dos alunos obtidas pela aplicação do teste em 2007/2008 comparativamente com o ano lectivo 2006/2007. Os alunos do primeiro

ano de aplicação do teste de escolha múltipla, estavam envolvidos num projecto de aprendizagem baseada em projectos interdisciplinares, designado PLE *(Project Led Education)*, que agregava todas as disciplinas com excepção da disciplina de Estatística I. Por várias vezes, os alunos reportaram ao professor da disciplina que o PLE lhes exigia demasiado tempo e esforço. Talvez seja esta a explicação para a diferença de médias entre os dois anos.

As abordagens TCT e MRI foram usadas de modo complementar; porém a aplicação dos MRI possibilitou a análise dos itens do teste e não do teste como um todo, contribuindo assim para um melhor entendimento das classificações dos alunos neste tipo de testes e permitindo a aplicação de testes de qualidade. A utilização dos MRI permite a criação de bancos de itens e, assim, os itens/testes aplicados para a aferição das aprendizagens permitirão reunir condições para a modelação de dados com vista à obtenção da métrica que permita a comparação efectiva dos resultados atingidos pelos alunos em anos lectivos diferentes.

## 3.2 Politómicos

### Dados e Resultados

Os dados em análise foram recolhidos através da aplicação da prova de aferição de Matemática do 4º ano de escolaridade do Ensino Básico (PAM4) (Anexo 4), no final do ano lectivo 2006/2007. Estes dados estão descritos no Anexo 3, secção 3.2. Realizaram a PAM4 um total de 108441 alunos. Os resultados obtidos são baseados na análise das respostas dos 760 alunos que pertencem à região da Cova da Beira (CB), e que realizaram a PAM. Foi considerada esta amostra, porque a utilização dos MRI garante a independência do grupo de examinandos a que é aplicado o instrumento e, consequentemente, os parâmetros dos itens e do factor latente são invariantes. Os dados da PAM4 foram modelados utilizando o modelo de resposta

ao item de Crédito Parcial Generalizado, apresentado no capítulo 1, na subsecção 1.5.5. As estimativas dos parâmetros dos itens e do factor latente, desempenho em Matemática, são obtidas pela aplicação do procedimento de estimação de máxima verosimilhança marginal, recorrendo ao algoritmo EM (Baker e Kim [11]). A estimação que recorre a este procedimento é efectuada no programa computacional Parscale (Muraki e Bock [90]).

Começamos por analisar as estimativas dos parâmetros dos itens baseadas no MRI. Para tal, convertemos a pontuação dos itens[1] em, no máximo, 4 categorias $(c_f, f = 1, ..., 4)$. A PAM4 é constituída por itens de 2, 3 e 4 categorias. Os itens 1, 2, 4, 6, 11, 12, 13, 15, 19, 22, 23, 25 e 27 são compostos por 2 categorias, os itens 5, 7, 10, 14, 16, 21 e 24 têm 3 categorias e os itens 3, 8, 9, 17, 18, 20 e 26 apresentam 4 categorias. A análise das frequências das categorias de resposta em cada um dos itens (tabela 3.8) indica que, em geral, a distribuição de frequências das categorias inferiores e/ou superiores é mais elevada do que nas categorias centrais, com excepção dos itens 9, 18 e 20. O mesmo podemos constatar pela análise das estimativas obtidas para a dificuldade em alcançar pontuação em cada uma das categorias.

As estatísticas descritivas das estimativas dos parâmetros dos itens calculados via MRI (tabela 3.9), bem como a sua representação gráfica (figura 3.4) mostram que a distribuição da estimativa do parâmetro de discriminação $(\hat{a})$ dos itens é assimétrica positiva, com média 0,541 e mediana 0,502, coeficiente de assimetria 0,618 e erro padrão associado 0,448. No que se refere à estimativa do parâmetro de dificuldade $(\hat{b})$ a distribuição tem média -1,263, mediana -0,817 e coeficiente de assimetria -1,495, com um erro padrão associado 0,448. Os 27 itens que compõem o teste são maioritariamente fáceis e de dificuldade média. O 1º quartil da distribuição é -1,858 e o 3º quartil é -0,467.

Na tabela 3.10 consta a classificação dos itens segundo as estimativas dos parâme-

[1]Pontuação atribuída pelo GAVE. A pontuação varia entre 0 e 6 pontos; por exemplo, os itens 8 e 18 têm 4 categorias de resposta, pontuadas como 0, 2, 4, 6 e 0, 1, 2, 6, respectivamente.

Tabela 3.8: Distribuição de frequências das categorias de resposta a cada item e dificuldades para alcançar cada uma das categorias da PAM4

| | | Categorias | | | | Dificuldade | | |
|---|---|---|---|---|---|---|---|---|
| Item | Questão | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_2$ | $c_3$ | $c_4$ |
| 1 | 1.1 | 35,3 | 64,7 | - | - | - | - | - |
| 2 | 1.2 | 37,0 | 63,0 | - | - | - | - | - |
| 3 | 2 | 40,7 | 10,5 | 18,6 | 30,3 | -1,757 | 1,153 | 0,604 |
| 4 | 3 | 8,7 | 91,3 | - | - | - | - | - |
| 5 | 4 | 28,9 | 4,3 | 66,7 | - | -1,939 | 1,939 | - |
| 6 | 5.1 | 16,3 | 83,7 | - | - | - | - | - |
| 7 | 5.2 | 16,7 | 1,6 | 81,7 | - | -6,173 | 6,173 | - |
| 8 | 5.3 | 48,6 | 12,6 | 2,2 | 36,6 | -1,451 | -2,450 | 3,901 |
| 9 | 6.1 | 5,8 | 5,0 | 1,8 | 87,4 | -2,500 | -5,110 | 7,610 |
| 10 | 6.2 | 45,5 | 0,3 | 54,2 | - | -9,996 | 9,996 | - |
| 11 | 7 | 21,3 | 78,7 | - | - | - | - | - |
| 12 | 8.1 | 8,3 | 91,7 | - | - | - | - | - |
| 13 | 8.2 | 29,7 | 70,3 | - | - | - | - | - |
| 14 | 9 | 3,0 | 2,2 | 94,7 | - | -3,979 | 3,979 | - |
| 15 | 10 | 20,0 | 80,0 | - | - | - | - | - |
| 16 | 11.1 | 25,3 | 0,4 | 74,3 | - | -5,327 | 5,327 | - |
| 17 | 11.2 | 31,4 | 7,4 | 3,0 | 58,2 | -1,892 | -1,521 | 3,413 |
| 18 | 12 | 2,9 | 0,8 | 20,7 | 75,7 | -2,064 | 2,452 | -0,388 |
| 19 | 13 | 16,2 | 83,8 | - | - | - | - | - |
| 20 | 14 | 19,9 | 20,0 | 12,6 | 47,5 | 0,159 | -0,869 | 0,710 |
| 21 | 15.1 | 31,6 | 1,4 | 67,0 | - | -5,340 | 5,340 | - |
| 22 | 15.2 | 52,8 | 47,2 | - | - | - | - | - |
| 23 | 16 | 32,6 | 67,4 | - | - | - | - | - |
| 24 | 17 | 13,8 | 2,9 | 83,3 | - | -3,832 | 3,832 | - |
| 25 | 18 | 17,2 | 82,8 | - | - | - | - | - |
| 26 | 19 | 29,7 | 11,2 | 16,3 | 42,8 | -1,005 | 0,349 | 0,656 |
| 27 | 20 | 45,4 | 54,6 | - | - | - | - | - |

Tabela 3.9: Estatísticas descritivas das estimativas dos parâmetros dos itens da PAM4

| | $\hat{a}$ | $\hat{b}$ |
|---|---|---|
| N | 27 | 27 |
| Média | 0,541 | -1,263 |
| Mediana | 0,502 | -0,817 |
| Desvio Padrão | 0,222 | 1,253 |
| Coeficiente de Assimetria | 0,618 | -1,495 |
| Erro Padrão da Assimetria | 0,448 | 0,448 |
| Mínimo | 0,219 | -5,159 |
| Máximo | 1,033 | 0,241 |
| Percentil 25 | 0,355 | -1,858 |
| Percentil 75 | 0,714 | -0,467 |

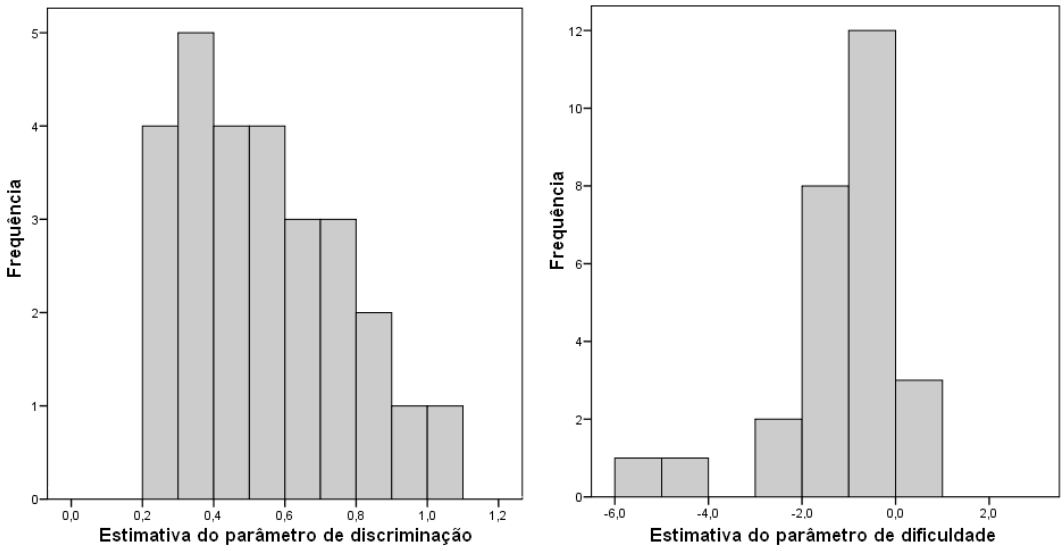



Figura 3.4: Histogramas das estimativas dos parâmetros de discriminação e dificuldade dos itens da PAM4

tros de discriminação e dificuldade encontrados via MRI.

A representação gráfica da dispersão dos itens, de acordo com os índices de discriminação e de dificuldade via MRI, está apresentada na figura 3.5.

Pela análise da tabela e do gráfico, é possível constatar a existência de apenas um item muito discriminativo (item 19) e a ausência de itens difíceis. Dos 13 itens

Tabela 3.10: Classificação dos itens da PAM4 face às estimativas dos parâmetros de discriminação e dificuldade

| **Discriminação** | **Limites** | **Itens** | **Número de Itens** | **Fr (%)** |
|---|---|---|---|---|
| Pouco Discriminativos (PD) | <0,5 | 3, 4, 7, 8, 9, 10, 14, 16, 17, 21, 22, 24, 26 | 13 | 48 |
| Discriminativos (D) | [0,5; 1[ | 1, 2, 5, 6, 11, 12, 13, 15, 18, 20, 23, 25, 27 | 13 | 48 |
| Muito Discriminativos (MD) | 1 | 19 | 1 | 4 |
| **Dificuldade** | **Limites** | **Itens** | **Número de Itens** | **Fr (%)** |
| Fácil (F) | [-3; -1[ | 4, 6, 7, 9, 11, 12, 14, 15, 18, 19, 24, 25 | 12 | 44,4 |
| Médio (M) | [-1; 1[ | 1, 2, 3, 5, 8, 10, 13, 16, 17, 20, 21, 22, 23, 26, 27 | 15 | 55,6 |
| Difícil (D) | [1; 3] | - | 0 | 0 |

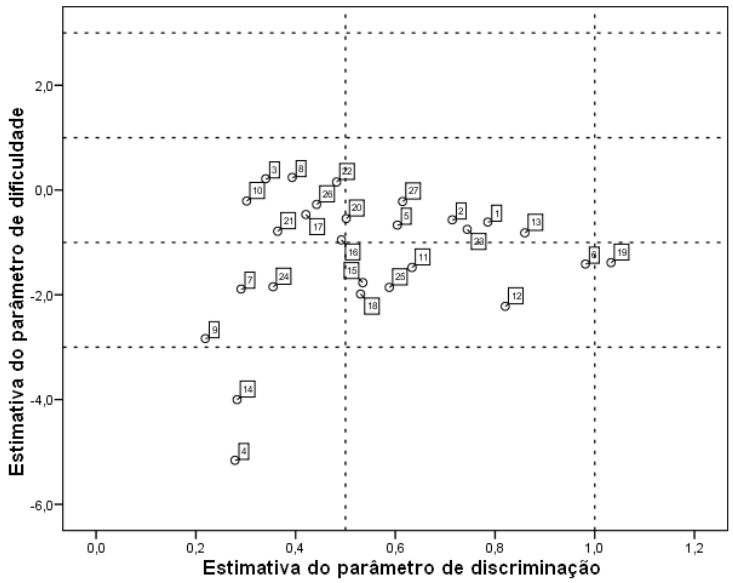



Figura 3.5: Gráfico de dispersão dos itens da PAM4, estimativa do parâmetro de discriminação por estimativa do parâmetro de dificuldade

com discriminação média, 6 itens são de nível fácil (6, 11, 12, 15, 18 e 25) e 7 itens são de nível de dificuldade média (1, 2, 5, 13, 20, 23 e 27). Dos restantes 13 itens, que apresentam baixa discriminação, 5 itens são fáceis (4, 7, 9, 14 e 24) e 8 itens são de dificuldade média (3, 8, 10, 16, 17, 21, 22 e 26).

Considerando as respostas dos alunos de Portugal Continental foi calculada a correlação de Pearson entre as estimativas dos parâmetros dos itens obtidas para a Cova da Beira e para Portugal Continental (tabela 3.11).

Tabela 3.11: Correlação entre as estimativas dos parâmetros dos itens obtidas para a região da Cova da Beira e para Portugal Continental

| Correlação de Pearson | PAM4 |
|---|---|
| $\hat{a}$ | 0,961 |
| $\hat{b}$ | 0,977 |

Os coeficientes de correlação de Pearson encontrados demonstram uma correlação forte positiva, entre as estimativas dos parâmetros encontrados para a região da Cova da Beira e as estimativas dos parâmetros obtidas, considerando toda a população.

A tabela 3.12 apresenta a classificação de cada item em termos das estimativas dos parâmetros de discriminação e de dificuldade obtidas para os dados da região da Cova da Beira e nacionais.

Em geral, constatamos que os resultados da classificação das estimativas dos parâmetros dos itens para a região da Cova da Beira e para Portugal Continental são semelhantes. Destacamos apenas que nos itens 6, 15, 16, 19 e 22 existe alguma discrepância na classificação da estimativa do parâmetro de discriminação entre a Cova da Beira e Portugal Continental.

De forma a complementar o estudo realizado, foi também efectuada a análise dos itens que compõem a prova através da abordagem baseada na TCT (Relatório 1: Provas de Aferição de Matemática, Português do 4º e 6º anos de escolaridade [98]). Os resultados apurados não são apresentados nesta secção, uma vez que não fazem parte do objectivo da aplicação de MRI a itens politómicos. No entanto, verificámos

Tabela 3.12: Classificação dos itens PAM4 face às estimativas dos parâmetros de discriminação e dificuldade para a Cova da Beira e Portugal Continental

| Item | Questão | $\hat{a}$ | | $\hat{b}$ | |
|---|---|---|---|---|---|
| | | Cova da Beira | Portugal Continental | Cova da Beira | Portugal Continental |
| 1 | 1.1 | D | D | M | M |
| 2 | 1.2 | D | D | M | M |
| 3 | 2 | PD | PD | M | M |
| 4 | 3 | PD | PD | F | F |
| 5 | 4 | D | D | M | M |
| 6 | 5.1 | D | MD | F | F |
| 7 | 5.2 | PD | PD | F | F |
| 8 | 5.3 | PD | PD | M | M |
| 9 | 6.1 | PD | PD | F | F |
| 10 | 6.2 | PD | PD | M | M |
| 11 | 7 | D | D | F | F |
| 12 | 8.1 | D | D | F | F |
| 13 | 8.2 | D | D | M | M |
| 14 | 9 | PD | PD | F* | F |
| 15 | 10 | D | PD | F | F |
| 16 | 11.1 | PD | D | M | M |
| 17 | 11.2 | PD | PD | M | M |
| 18 | 12 | D | D | F | F |
| 19 | 13 | MD | D | F | F |
| 20 | 14 | D | D | M | M |
| 21 | 15.1 | PD | PD | M | M |
| 22 | 15.2 | PD | D | M | M |
| 23 | 16 | D | D | M | M |
| 24 | 17 | PD | PD | F | F |
| 25 | 18 | D | D | F | F |
| 26 | 19 | PD | PD | M | M |
| 27 | 20 | D | D | M | M |

que a análise conjunta dos parâmetros que caracterizam os itens segundo a TCT e a MRI sugerem que os itens 4, 9 e 14 são itens pouco discriminativos. O item 19 é o único item comum às duas análises que é muito discriminativo. Relativamente ao nível de dificuldade, denota-se que, aproximadamente, 44% dos itens que compõem a prova são fáceis, sendo que os itens 4 e 14 são extremamente fáceis. Assim, podemos

concluir que a análise realizada através das duas abordagens conduz a resultados corroborantes.

A título de ilustração, apresentamos seguidamente a curva característica e a função de informação para três dos itens da prova com 2, 3 e 4 categorias de resposta, respectivamente. A curva característica e a função de informação dos restantes itens da prova podem ser consultados no anexo 2 do Relatório 1: Provas de Aferição de Matemática, Português do 4º e 6º anos de escolaridade [98]. Daqui em diante, designaremos a estimativa do factor latente a Matemática por desempenho em Matemática.

O item 19 é dicotómico e foi respondido correctamente por 83,8% dos alunos, como pode ser verificado na tabela 3.8. Este item é de baixa dificuldade, tanto pela classificação da estimativa do parâmetro de dificuldade, como pela observação do eixo horizontal da curva característica do item. É um item muito discriminativo, como se pode constatar pela classificação da estimativa deste parâmetro apresentada na tabela . A função de informação do item mostra que este item apresenta melhor informação para alunos com desempenho entre -2,5 e -0,5 (figura 3.6).

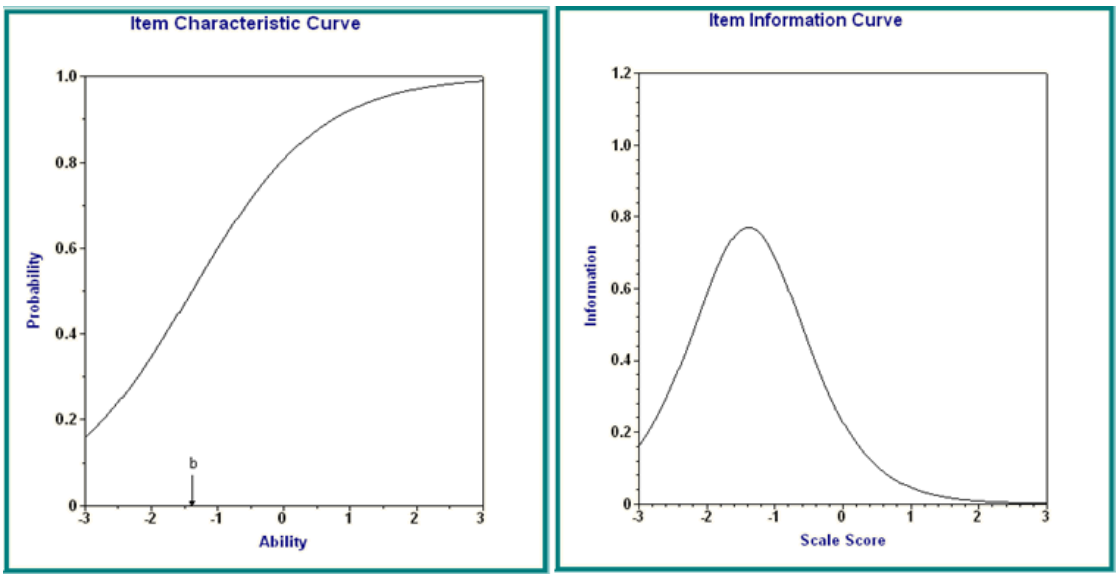



Figura 3.6: Curva característica e função de informação do item 19

O item 5 apresenta 3 categorias de resposta e 66,7% dos alunos acertaram completamente a questão, 28,9% erraram e 4,3% obtiveram acerto parcial no item. Nesse sentido, e baseando-nos também na classificação da estimativa do parâmetro de dificuldade, concluimos que o item apresenta dificuldade média. No que diz respeito à estimativa do parâmetro de discriminação, podemos verificar que este item é discriminativo. A curva característica deste item é constituída por 3 curvas relativas às categorias de resposta do item. Observamos que a categoria de resposta 1 tem maior probabilidade para níveis de desempenho mais baixos, a categoria 2 apresenta maior probabilidade para níveis de desempenho intermédios, embora inferior às outras categorias, e que a categoria 3 mostra maior probabilidade para níveis de desempenho mais elevados. A função de informação reflecte uma boa contribuição para a prova essencialmente para níveis de desempenho entre -2 e 0,5 (figura 3.7).

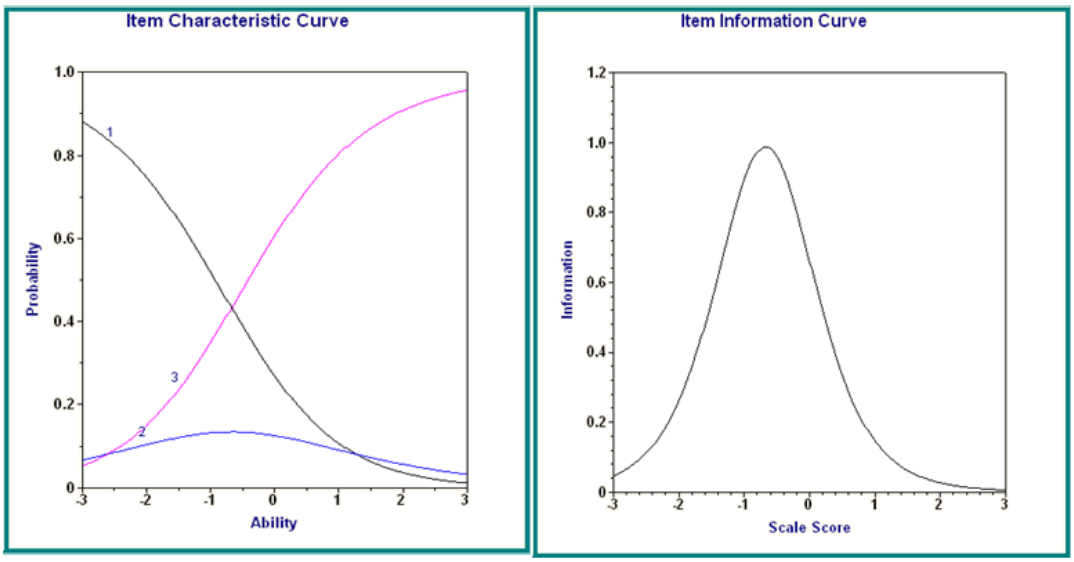



Figura 3.7: Curva característica e função de informação do item 5

No item 20, 47,5% dos alunos acertaram o item, 12,6% dos alunos obtiveram acerto parcial na categoria 3, 20% obtiveram acerto parcial na categoria 2 e 19,9% erraram o item. Este item é discriminativo e de dificuldade média. A curva característica deste item apresenta 4 curvas que indicam que a probabilidade de seleccionar

a categoria 1 relativamente às categorias 2 e 3 é maior em níveis de desempenho mais baixos do que nos mais altos. Adicionalmente, para níveis de desempenho mais altos a probabilidade de obter pontuação na categoria 4 é superior à de obter pontuação nas categorias 2 e 3. Este item contribui para a prova, essencialmente, para níveis de desempenho entre -2 e 1 (figura 3.8).

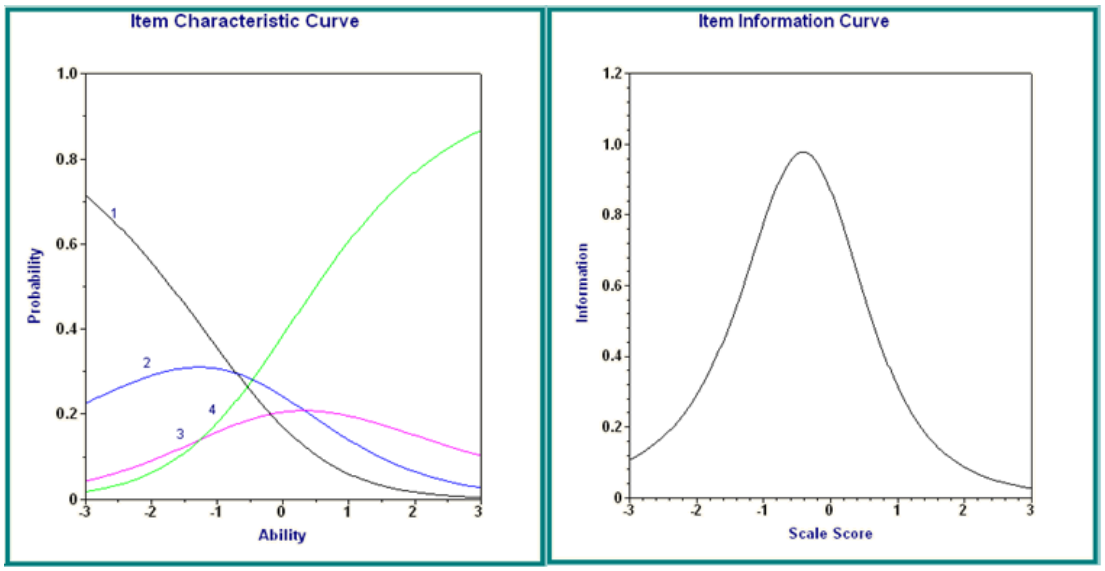



Figura 3.8: Curva característica e função de informação do item 20

A análise da função de informação do teste (Muraki [89]) indica que os dados recolhidos através deste instrumento contêm "grande"' poder informativo relativamente ao factor latente de alunos com nível de conhecimento compreendido entre -2 e 0,5 e que o erro padrão da medida mostra elevada imprecisão dos resultados obtidos no que se refere à aferição da aprendizagem de alunos com nível de conhecimento no extremo superior da escala (figura 3.9).

No que se refere à identificação das perguntas problema, o processo de análise dos itens baseado MRI permite identificar dois itens muito fáceis, com baixo poder discriminativo - os itens 4 e 14. Ambos os itens apresentam elevada percentagem de acerto (91% e 95%, respectivamente). O item 4 (corresponde à questão 3 da Parte A da prova) é uma questão de escolha múltipla sobre a identificação de sólidos, pelo

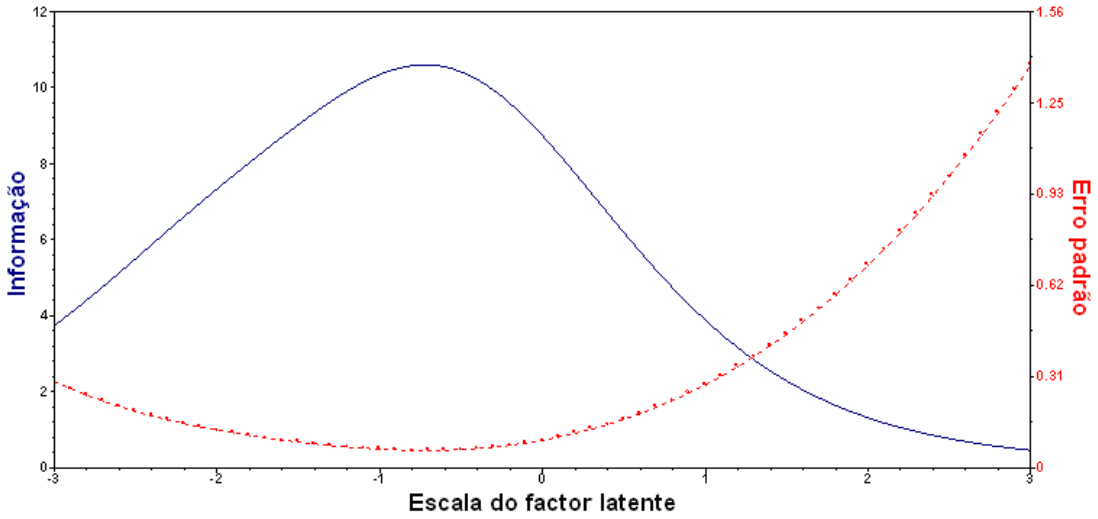



Figura 3.9: Função de informação do teste e erro padrão da PAM4

que pertencem ao conteúdo da Forma e Espaço. O item 14 (equivale à questão 9 da Parte B da prova) refere-se a um procedimento simples de cálculo de uma adição.

Procedemos, de seguida, à quantificação do impacto destes dois itens nos resultados escolares da PAM4, comparando as classificações obtidas sem as perguntas em causa. Calculámos o resultado final na prova retirando os itens 4 e 14, de modo a avaliar o seu impacto nos resultados. Nesse sentido, foi obtida a classificação dos alunos decorrente da aplicação do MRI para os 27 itens ($class - PAM4$) e retirando os dois itens ($class - PAM4 - 25itens$).

A tabela 3.13 e o gráfico de dispersão entre os resultados obtidos com todos os itens e retirando os itens 4 e 14, mostram que a correlação é quase perfeita (figura 3.10).

Tabela 3.13: Correlação de Pearson da classificação considerando 27 e 25 itens

| Correlação de Pearson | $class - PAM4 - 25itens$ |
|---|---|
| $class - PAM4$ | 0,999 |

A análise do ajuste do modelo de crédito parcial generalizado aos dados, feita

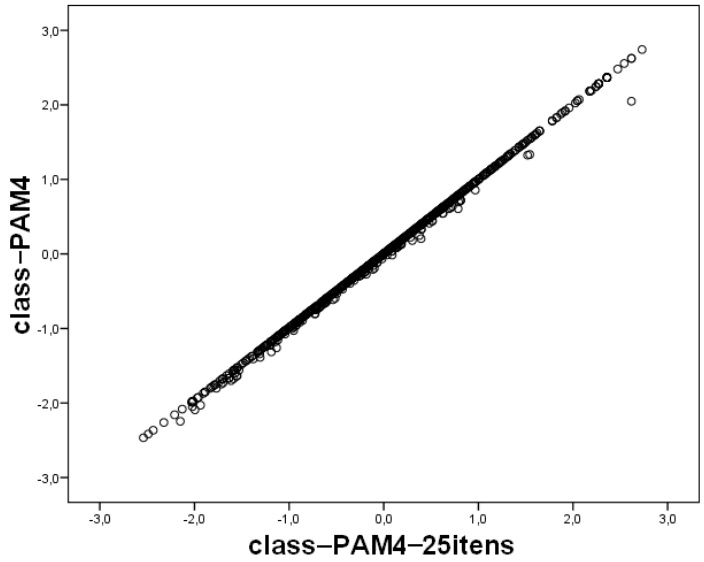



Figura 3.10: Diagrama de dispersão das classificações na PAM4 considerando 27 e 25 itens

com base na estatística de teste do qui-quadrado ($\chi^2$), indica ajuste global do modelo aos dados (tabela 3.14). A estatística é dada pela soma dos valores de $\chi^2$ para todos os itens e dos graus de liberdade respectivos, $\chi^2$=369,212 (g.l.=341, p-valor=0,141).

Tabela 3.14: Estatísticas de ajuste

| Item | Questão | $\chi^2$ | g.l. | p-valor | Item | Questão | $\chi^2$ | g.l. | p-valor |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.1 | 22,090 | 11 | 0,024 | 15 | 10 | 5,306 | 10 | 0,870 |
| 2 | 1.2 | 14,599 | 11 | 0,201 | 16 | 11.1 | 5,848 | 10 | 0,829 |
| 3 | 2 | 31,922 | 30 | 0,371 | 17 | 11.2 | 7,207 | 18 | 0,988 |
| 4 | 3 | 2,735 | 6 | 0,843 | 18 | 12 | 19,784 | 13 | 0,100 |
| 5 | 4 | 10,478 | 10 | 0,400 | 19 | 13 | 9,0347 | 8 | 0,339 |
| 6 | 5.1 | 11,571 | 8 | 0,171 | 20 | 14 | 27,309 | 27 | 0,447 |
| 7 | 5.2 | 18,860 | 10 | 0,042 | 21 | 15.1 | 15,795 | 11 | 0,148 |
| 8 | 5.3 | 25,759 | 21 | 0,215 | 22 | 15.2 | 20,326 | 14 | 0,120 |
| 9 | 6.1 | 6,191 | 12 | 0,906 | 23 | 16 | 17,990 | 11 | 0,081 |
| 10 | 6.2 | 8,574 | 13 | 0,805 | 24 | 17 | 11,607 | 10 | 0,312 |
| 11 | 7 | 9,196 | 10 | 0,514 | 25 | 18 | 7,467 | 10 | 0,682 |
| 12 | 8.1 | 1,517 | 6 | 0,957 | 26 | 19 | 33,696 | 29 | 0,250 |
| 13 | 8.2 | 4,459 | 10 | 0,924 | 27 | 20 | 19,891 | 12 | 0,069 |
| 14 | 9 | 0,000 | 0 | 0,000 | - | - | - | - | - |

### Discussão

Dada a natureza dos dados, politómicos, e com o intuito de extrair maior informação dos itens, modelámos os dados pelo MRI de Crédito Parcial Generalizado. Identificámos os níveis de desempenho para os quais a prova foi particularmente informativa, através da função de informação do teste. Verificámos que a PAM4 apresentou estimativas médias do parâmetro de discriminação de 0,54 e do parâmetro de dificuldade de -1,26. As estimativas encontradas indicaram que esta prova é constituída, essencialmente, por itens de discriminação média e baixa e de nível de dificuldade médio e baixo. Devem ser revistos ou retirados os itens 4 e 14 que se apresentaram como pouco discriminativos e muito fáceis. A função de informação do teste e o erro padrão da medida permitiram concluir que esta prova contém grande poder informativo relativamente ao factor latente de alunos com nível de desempenho baixo e médio e indica elevada imprecisão dos resultados obtidos no que se refere à aferição da aprendizagem de alunos com nível de desempenho no extremo superior da escala. Em geral, o ajuste global do modelo aos dados (com base na estatística de teste qui-quadrado) mostrou-se adequado.

A utilização de MRI nesta aplicação permitiu verificar que os parâmetros dos itens da prova não dependem da amostra de alunos utilizada, o que foi comprovado pelos resultados invariantes obtidos para a amostra de alunos da Cova da Beira e para a população de Portugal Continental. Outra vantagem da utilização de MRI é a possibilidade de comparação entre alunos da mesma população que tenham sido submetidos a provas totalmente diferentes. Este tipo de análise será realizada na secção 3.5 deste capítulo.

Os resultados obtidos ressaltam, ainda, a importância da utilização destas técnicas na análise das propriedades dos instrumentos e dos itens em avaliação educacional, assim como o seu potencial na construção e validação de instrumentos.

## 3.3 Grupos múltiplos

### Dados e Resultados

Os dados referem-se a uma amostra aleatória de 1930 alunos dos 1º, 2º e 3º anos de escolaridade do Ensino Básico da região da Cova da Beira e foram recolhidos no âmbito do projecto 3EM (Anexo 3, secção 3.3). Para cada teste consideram-se as aplicações do início e do final do ano lectivo. A distribuição do número de alunos por ano lectivo e por aplicação é apresentada na tabela 3.15.

Tabela 3.15: Distribuição dos alunos por ano lectivo e por aplicação

| Ano lectivo | Aplicação | Frequência absoluta | Percentagem |
|---|---|---|---|
| 1º ano | Início | 309 | 16,0 |
| | Final | 304 | 15,8 |
| 2º ano | Início | 350 | 18,1 |
| | Final | 336 | 17,4 |
| 3º ano | Início | 315 | 16,3 |
| | Final | 316 | 16,4 |
| | Total | 1930 | 100,0 |

A distribuição de alunos por ano lectivo e por aplicação indica que não existem grandes diferenças na representatividade de cada um deles para a amostra.

Para estabelecer uma métrica comum entre os resultados dos examinandos em instrumentos diferentes, é necessário que existam itens comuns aos instrumentos, denominados itens âncora. Os testes têm itens comuns tanto no mesmo ano de escolaridade como em anos de escolaridade subsequentes, o que permitiu a criação de uma métrica comum. O número de itens comuns em cada teste é apresentado na tabela 3.16.

Constata-se que existe um maior número de itens comuns entre as aplicações do ínicio e do final de cada ano de escolaridade do que entre anos de escolaridade diferentes. A percentagem de itens comuns indicada para fazer equalização é de 20%, no mínimo, o que se verifica neste caso.

Tabela 3.16: Número de itens comuns entre cada teste

| | | 1º ano | | 2º ano | | 3º ano | |
|---|---|---|---|---|---|---|---|
| | | Início | Final | Início | Final | Início | Final |
| 1º ano | Início | 30 | 19 | 8 | 7 | 0 | 0 |
| | Final | 19 | 30 | 9 | 9 | 0 | 0 |
| 2º ano | Início | 8 | 9 | 30 | 24 | 9 | 9 |
| | Final | 7 | 9 | 24 | 30 | 9 | 9 |
| 3º ano | Início | 0 | 0 | 9 | 9 | 30 | 23 |
| | Final | 0 | 0 | 9 | 9 | 23 | 30 |

Uma análise exploratória dos dados foi inicialmente realizada com o objectivo de analisar os itens que compunham cada teste. Para isso, recorremos a algumas estatísticas da TCT, nomeadamente, ao índice de dificuldade e à correlação bisserial calculados para os itens de cada teste. Optámos por retirar os itens que tinham correlação bisserial negativa. O teste do 1º ano - início passou a ser composto por 29 itens, o do 1º ano - final ficou com um total de 28 itens; o do 2º ano - início passou a ter 22 itens e o do 2º ano-final ficou constituído por 21 itens. Os testes do 3º ano, tanto no início como no final do ano, ficaram com 22 itens. É de realçar que alguns dos itens retirados eram comuns a 2 ou mais testes. Nesse sentido, foram considerados para a análise dos resultados um total de 144 itens.

Os parâmetros dos itens e o factor latente foram estimados conjuntamente, considerando todos os itens, anos de escolaridade e aplicações, pelo software Bilog-MG (Zimowski *et al.* [118]). O procedimento de estimação adoptado neste software foi o da máxima verosimilhança, o que permite a estimação conjunta dos parâmetros em várias amostras não-equivalentes. O *design* de todos os testes com os itens que compunham cada um deles foi inserido no software.

A utilização do modelo para grupos múltiplos permitiu a obtenção das estimativas dos parâmetros dos itens e do factor latente, dos alunos dos 1º, 2º e 3º anos de escolaridade que realizaram os testes, numa métrica comum. Esta aplicação faz parte do procedimento adoptado para realizar a análise da dimensionalidade dos testes de Matemática. A análise da dimensionalidade dos testes foi efectuada pela

aplicação do modelo multidimensional logístico de 3 parâmetros. Os procedimentos estatísticos usados são exemplificados no capítulo 4, secção 4.1. É de notar que a versão do software usado para realizar esta análise, Testfact 2.13 (Wilson, Wood e Gibbons [115]), apenas possibilita a estimação dos parâmetros do modelo multidimensional logístico de 2 parâmetros. Contudo, este software permite que o parâmetro de acerto ao acaso seja estimado noutro software e incorporado na sintaxe do Testfact (Wilson, Wood e Gibbons [115]). Costa *et al.* [26] evidenciaram que, considerar o parâmetro de probabilidade de acerto ao acaso aumenta a informação, nomeadamente, em termos de percentagem de variância explicada para cada factor, pelo que deve ser considerado na análise dos resultados. Na tabela 3.17 são apresentadas algumas estatísticas descritivas obtidas para as estimativas do parâmetro de probabilidade de acerto ao acaso.

Tabela 3.17: Estatísticas descritivas das estimativas do parâmetro de probabilidade de acerto ao acaso

| | $\hat{c}$ |
|---|---|
| Média | 0,155 |
| Mediana | 0,144 |
| Desvio Padrão | 0,070 |
| Mínimo | 0,020 |
| Máximo | 0,318 |
| Percentil 25 | 0,105 |
| Percentil 75 | 0,194 |

A análise da tabela permite verificar que a distribuição das estimativas do parâmetro de probabilidade de acerto ao acaso, para os itens dos testes de todos os anos de escolaridade, tem média 0,155 e mediana 0,144. O 1º quartil da distribuição é 0,105, o que quer dizer que 25% dos itens dos testes têm probabilidade de acerto ao acaso inferior a 0,105 e o 3º quartil é 0,194.

Assim, nesta aplicação, as estimativas do parâmetro de probabilidade de acerto ao acaso, obtidas pela aplicação do modelo para grupos múltiplos, foram inseridas no Testfact 2.13 (Wilson, Wood e Gibbons [115]), o que pemitiu a utilização

do modelo logístico multidimensional de 3 parâmetros e a consequente análise da dimensionalidade dos testes de Matemática.

### Discussão

A utilização de itens comuns e o uso de modelos para grupos múltiplos permitiram a obtenção de uma métrica comum para as estimativas dos parâmetros dos itens e do factor latente em todos os anos lectivos e aplicações.

A obtenção de uma estimativa única dos parâmetros do modelo, evita a propagação de erros em termos de equalização comparativamente com a estimação dos parâmetros aplicação a aplicação, ou ano a ano. O recurso a este procedimento permite fazer comparações e acompanhar a evolução do ensino/aprendizagem ao longo dos anos. A utilização desta classe de modelos permite que a equalização seja feita simultaneamente.

## 3.4 Equalização

### Dados e Resultados

Os dados foram obtidos no âmbito do projecto 3EM (Anexo 3 - secção 3.3) e referem-se aos testes 3EMat do 1º ano de escolaridade, aplicação do início do ano lectivo 2005/2006 e do 2º ano de escolaridade, aplicação do início do ano lectivo 2006/2007. Foram observados 309 alunos do 1º ano e 350 alunos do 2º ano, existindo 286 alunos que responderam aos dois testes. Ambos os testes são compostos por 30 itens. No teste do 2º ano existiam itens de 1º ano conforme descrito na subsecção 1.9.5 do capítulo 1, no plano de recolha de dados iii). Daqui em diante, utilizar-se-á "classificação" para designar a estimativa do factor latente/desempenho a Matemática.

A comparação das classificações dos alunos nos dois anos de escolaridade só é

possível ser realizada se todos os valores dos parâmetros dos itens e do factor latente estiverem na mesma escala de medida. Nesse sentido, escalas de desempenho a Matemática obtidas separadamente para diferentes anos lectivos não são passíveis de comparação, surgindo daí, a necessidade da utilização de métodos de equalização.

O modelo de resposta ao item logístico de 2 parâmetros (equação 1.4.2) foi utilizado para a obtenção das escalas de desempenho a Matemática. A escala de desempenho a Matemática foi obtida separadamente para o 1º ano de escolaridade (classificação 1) de forma a ter média zero e variância um.

Para a criação da escala comum utilizámos o método de equalização linear via itens comuns, que já foi descrito na subsecção 1.9.6 do capítulo 1.

A tabela 3.18 apresenta as estimativas do parâmetro de discriminação *($\hat{a}$_1 e $\hat{a}$_2)* e de dificuldade ($\hat{b}$_1 e $\hat{b}$_2) dos 7 itens âncora, nos testes do 1º e 2º anos, respectivamente.

Tabela 3.18: Equalização via itens comuns - Estimativas dos parâmetros de discriminação e de dificuldade dos itens âncora

| Item | $\hat{a}$_1 | $\hat{b}$_1 | $\hat{a}$_2 | $\hat{b}$_2 |
|---|---|---|---|---|
| 1 | 0,646 | 3,024 | 0,972 | 0,099 |
| 2 | 0,409 | -0,219 | 0,806 | -0,380 |
| 3 | 0,497 | 0,126 | 0,346 | -2,687 |
| 4 | 0,486 | 3,930 | 0,828 | -0,115 |
| 5 | 0,491 | 0,307 | 0,554 | 0,168 |
| 6 | 0,530 | 2,318 | 0,866 | 0,220 |
| 7 | 0,751 | 2,408 | 0,848 | 0,585 |

A análise das estimativas dos parâmetros dos itens âncora permitem-nos identificar possíveis observações extremas. Através da tabela 3.18 verificamos que do 1º ano para o 2º ano, as estimativas do parâmetro de discriminação aumentam, com excepção do item 3, e que as estimativas do parâmetro de dificuldade diminuem. Constata-se ainda que, no teste do 1º ano o item 2 apresenta as estimativas mais baixas dos seus parâmetros, o que não se verifica no teste do 2º ano. Para ana-

lisar o comportamento do item 2 comparámos, em ambos os testes, as suas curvas características e as respectivas funções de informação. Adicionalmente, baseado no procedimento descrito por Kolen e Brennan [71], analisámos os diagramas de dispersão das estimativas do parâmetro de discriminação e das estimativas do parâmetro de dificuldade dos itens âncora. Retirando o item 2, foi obtido um melhor ajuste à recta de regressão das estimativas dos parâmetros dos itens. O coeficiente de correlação de Pearson para os 7 itens âncora e para os 6 itens âncora no diagrama de dispersão das estimativas do parâmetro de discriminação passou de 0,151 para 0,284, respectivamente, e, no caso das estimativas do parâmetro de dificuldade passou de 0,235 para 0,302, respectivamente. Estes procedimentos permitem verificar que o item 2 é uma observação extrema.

A equalização via itens comuns (Kolen e Brennan [71]) foi efectuada para os 7 e os 6 itens âncora, dado o número reduzido de itens comuns. Utilizámos o procedimento Média-Desvio (Marco [81]) e o procedimento Média-Média (Loyd e Hoover [79]), considerando como população de referência os alunos que responderam ao teste do 1º ano. No procedimento Média-Desvio, para os 7 itens âncora $A$=1,480 (obtida por 1.9.6) e $B$=2,145 (obtida por 1.9.9) e para os 6 itens âncora $A$=1,262 e $B$=2,383. No procedimento Média-Média, para os 7 itens âncora $A$=1,371 (obtida por 1.9.7) e $B$=2,112 (obtida por 1.9.9) e para os 6 itens âncora $A$=1,298 e $B$=2,393.

As tabelas 3.19 e 3.20 contêm algumas estatísticas descritivas das classificações obtidas pela aplicação do método de equalização via itens comuns. Nestas tabelas, classificação 2.1 e classificação 2.2 representam as classificações obtidas pelo procedimento Média-Desvio, e, classificação 3.1 e classificação 3.2 representam as classificações obtidas pelo procedimento Média-Média, ambos para os 7 e os 6 itens âncora, respectivamente.

A análise das tabelas permite verificar que do 1º ano para o 2º ano houve uma melhoria da média e dos quartis em todas as classificações obtidas.

Com vista à comparação do desempenho a Matemática do 1º para o 2º ano,

Tabela 3.19: Estatísticas descritivas das classificações obtidas na equalização via itens comuns (1ª parte)

| Estatísticas | classificação 1 | classificação 2.1 | classificação 2.2 |
|---|---|---|---|
| N | 286 | 286 | 286 |
| Média | 0,000 | 2,277 | 2,494 |
| Mediana | -0,121 | 2,256 | 2,477 |
| Desvio padrão | 0,995 | 1,615 | 1,377 |
| Mínimo | -2,246 | -2,960 | -1,970 |
| Máximo | 3,390 | 7,290 | 6,770 |
| Percentil 25 | -0,651 | 1,208 | 1,583 |
| Percentil 75 | 0,579 | 3,119 | 3,281 |

Tabela 3.20: Estatísticas descritivas das classificações obtidas na equalização via itens comuns (2ª parte)

| Estatísticas | classificação 1 | classificação 3.1 | classificação 3.2 |
|---|---|---|---|
| N | 286 | 286 | 286 |
| Média | 0,000 | 2,248 | 2,522 |
| Mediana | -0,121 | 2,210 | 2,486 |
| Desvio padrão | 0,995 | 1,493 | 1,413 |
| Mínimo | -2,246 | -2,490 | -1,960 |
| Máximo | 3,390 | 6,880 | 6,910 |
| Percentil 25 | -0,651 | 1,260 | 1,586 |
| Percentil 75 | 0,579 | 3,098 | 3,327 |

aplicámos o teste de hipóteses para amostras emparelhadas. Realizámos o teste de hipóteses e os resultados obtidos corroboram a hipótese de que a classificação dos alunos do 2º ano é estatísticamente superior à classificação do 1º ano. Concretamente, para todos os procedimentos adoptados o valor da estatística de teste e o respectivo valor de prova é t=20,655 (p<0,01), t=25,297 (p<0,01), t=28,435 (p<0,01), t=33,375 (p<0,01).

As tabelas 3.21 e 3.22 contêm algumas estatísticas descritivas da função de informação do teste e do erro padrão de medida, calculados nos itens âncora, para as classificações obtidas nos diferentes procedimentos. Teste 1.1 e teste 1.2 representam os valores obtidos no teste do 1º ano, para os 7 itens âncora e para os 6 itens âncora,

respectivamente.

Tabela 3.21: Estatísticas descritivas da função de informação do teste

| Estatísticas | Teste 1.1 | Teste 1.2 | Teste 2.1 | Teste 2.2 | Teste 3.1 | Teste 3.2 |
|---|---|---|---|---|---|---|
| Máximo | 0,483 | 0,471 | 0,490 | 0,757 | 0,571 | 0,716 |
| Mínimo | 0,247 | 0,198 | 0,148 | 0,116 | 0,172 | 0,110 |
| Média | 0,383 | 0,330 | 0,420 | 0,490 | 0,490 | 0,463 |

A tabela anterior permite verificar que é nos testes 2.2 e 3.1 que a média da função de informação do teste é mais elevada, comparativamente com os restantes testes, e, que é no teste 1.2 que a função de informação do teste contribui com menor informação.

Tabela 3.22: Estatísticas descritivas do erro padrão de medida

| Estatísticas | Teste 1.1 | Teste 1.2 | Teste 2.1 | Teste 2.2 | Teste 3.1 | Teste 3.2 |
|---|---|---|---|---|---|---|
| Máximo | 2,011 | 2,250 | 2,603 | 2,938 | 2,411 | 3,021 |
| Mínimo | 1,438 | 1,467 | 1,428 | 1,149 | 1,323 | 1,182 |
| Média | 1,635 | 1,769 | 1,564 | 1,520 | 1,449 | 1,563 |

O erro padrão de medida apresenta, em média, o valor mais baixo no teste 3.1 e o valor mais alto é o do teste 1.2. Assim, verifica-se que o procedimento Média-Média para os 7 itens âncora (teste 3.1) é o que tem, em média, menor erro padrão da medida.

Com vista à comparação dos resultados da equalização linear, para os procedimentos Média-Média e Média-Desvio calcularam-se o *Hdiff* (Haebara [56]) e o *Sldiff* (Stocking e Lord [108]), considerando os valores fixos da classificação (0; 1; -1; 0,5 e -0,5) para os 7 (2.1 e 3.1) e os 6 (2.2 e 3.2) itens âncora. Os valores obtidos encontram-se na tabela 3.23.

Constata-se que o *Hdiff* para os diferentes valores da classificação é menor no procedimento Média-Média para os 6 itens âncora, e no *SLdiff* verifica-se que, em geral, os valores são menores no mesmo procedimento, mas para os 7 itens âncora. Estes resultados vêm mostrar que o procedimento Média-Média apresenta valores

Tabela 3.23: Critérios *Hdiff* e *Sldiff* para comparar os procedimentos Média-Média e Média-Desvio

| Critério | Classificação | 2.1 | 2.2 | 3.1 | 3.2 |
|---|---|---|---|---|---|
| | 0 | 0,252 | 0,303 | 0,236 | 0,025 |
| | 1 | 0,296 | 0,379 | 0,294 | 0,031 |
| Hdiff | -1 | 0,195 | 0,193 | 0,194 | 0,015 |
| | 0,5 | 0,277 | 0,353 | 0,261 | 0,030 |
| | -0,5 | 0,223 | 0,246 | 0,214 | 0,019 |
| | 0 | 0,143 | 0,411 | 0,087 | 0,371 |
| | 1 | 0,181 | 0,677 | 0,194 | 0,623 |
| SLdiff | -1 | 0,024 | 0,104 | 0,007 | 0,090 |
| | 0,5 | 0,191 | 0,588 | 0,129 | 0,536 |
| | -0,5 | 0,075 | 0,233 | 0,038 | 0,207 |

mais estáveis e, consequentemente, é o mais adequado para efectuar a equalização, tal como acontece em Baker e Al-Karni [10].

### Discussão

Com vista a obter uma escala vertical de desempenho escolar a Matemática no Ensino Básico, foi aplicado o método de equalização via itens comuns. Para a equalização via itens comuns foram aplicados o procedimento Média-Desvio e o procedimento Média-Média. Em ambos os procedimentos houve uma melhoria das classificações do 1º ano de escolaridade para o 2º ano de escolaridade.

O erro padrão de medida obtido para as classificações nos itens âncora foi menor no procedimento Média-Média. O trabalho empírico desenvolvido corroborou a evidência registada na literatura de que o procedimento Média-Média produz estimativas dos parâmetros mais estáveis.

No âmbito do projecto 3EM, o procedimento adoptado nesta subsecção poderá ser utilizado para a obtenção de uma escala vertical única padronizada do desempenho a Matemática considerando diferentes anos escolaridade do Ensino Básico. Em termos de avaliação educacional, este tipo de procedimentos pode ser aplicado

a outros instrumentos de aferição de aprendizagens.

Instrumentos de aferição de aprendizagens diferentes podem não ter a mesma dificuldade ou a mesma distribuição para o factor latente, o que pode indicar que a interpretação das classificações não tenham o mesmo significado. Assim, a utilização de MRI permite aferir o desempenho em instrumentos e, através da aplicação de procedimentos de equalização, possibilita a comparação de classificações dos examinandos em instrumentos de aferição de aprendizagens diferentes ou em instrumentos que são administrados em fases/ocasiões diferentes.

## 3.5 *Linking*

### Dados e Resultados

Esta aplicação foi desenvolvida no contexto do projecto Melhoria da Qualidade dos Instrumentos e Escalas de Aferição dos Resultados Escolares (Anexo 3 - secção 3.2). No âmbito desta aplicação efectua-se o *linking*/ligação entre o teste 3EMat e a prova de aferição (PAM6), considerando a disciplina de Matemática do 6º ano de escolaridade no ano lectivo 2006/2007. Neste ano lectivo, a data de aplicação do teste 3EMat difere da data de aplicação da PAM6 em quinze dias. Assumindo o pressuposto que ambos os instrumentos aferem as mesmas aprendizagens, o *linking* entre as duas escalas foi concretizado através dos alunos que realizaram tanto a PAM6 como o teste 3EMat, num total de 303.

Na apreciação da validade externa da PAM6 utilizámos o teste 3EMat. Ambas as escalas de desempenho a Matemática foram obtidas separadamente através da utilização de MRI. O procedimento de estimação usado é o MVM, recorrendo ao algoritmo EM (Baker e Kim [11]). A estimação que recorre a este procedimento é efectuada no programa computacional Parscale (Muraki e Bock [90]).

A ligação entre as escalas subjacentes aos intrumentos PAM6 e teste 3EMat foi

realizada via população pela aplicação do método linear (Kolen e Brennan [71]). Neste método postula-se que as estimativas dos parâmetros associados ao factor latente, desempenho em Matemática, obtidas na PAM6 e no teste 3EMat estão linearmente relacionadas.

A comparação das classificações dos alunos em ambos os instrumentos só é possível ser realizada se todos os valores dos parâmetros dos itens e do factor latente estiverem na mesma escala de medida. Considerámos como grupo de referência os alunos que efectuaram a PAM6. A escala de desempenho a Matemática para esta prova foi obtida de forma a ter média zero e variância um (class-PAM6). Separadamente, obtiveram-se as classificações dos alunos no teste 3EMat (class-3EMat). A aplicação do procedimento de ligação permitiu a obtenção de uma escala de desempenho para o teste 3EMat, denominada class-3EMat-link, comparável com a escala do PAM6. Adicionalmente, foram calibrados conjuntamente os itens de ambos os instrumentos, obtendo-se uma escala conjunta de desempenho a Matemática (class-PAM6-3EMat).

O software utilizado para a obtenção das estimativas do factor latente eliminou um aluno (aluno nº 121), uma vez que este teve 100% de acerto em ambos os instrumentos. Assim, daqui em diante, são considerados apenas os restantes 302 alunos da amostra.

A distribuição das classificações, bem como a diferença entre as classificações da PAM6 e do teste 3EMat, estão representadas nos histogramas da figura 3.11.

Seguidamente, são apresentadas as estatísticas descritivas das classificações (tabela 3.24), o diagrama de extremos e quartis (figura 3.12) e o gráfico dos intervalos de confiança para as médias das classificações da PAM6 e do teste 3EMat (figura 3.13).

Quer pela análise dos gráficos dos intervalos de confiança para as médias das classificações, quer pela aplicação do teste de Mann-Whitney (as classificações não verificam o pressuposto da normalidade), cujo valor de prova é $p = 0,857$, para $\alpha =$

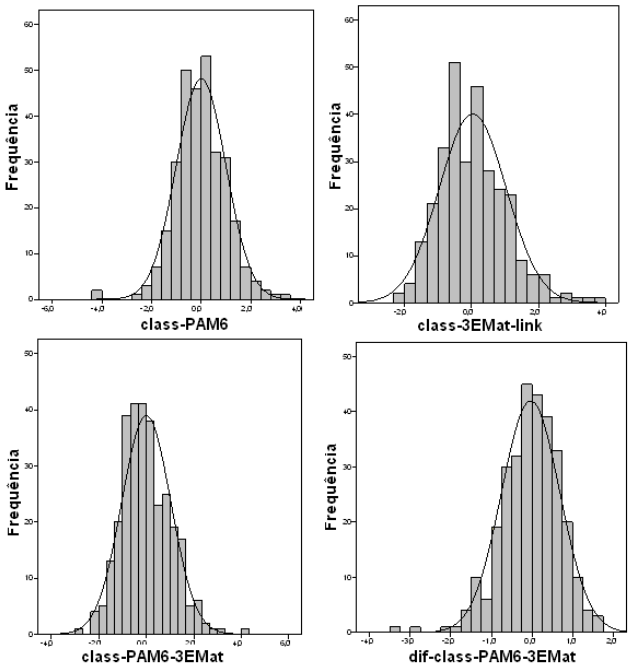



Figura 3.11: Histogramas das classificações dos instrumentos e das diferenças de classificação entre os instrumentos

Tabela 3.24: Estatísticas descritivas das classificações

| | class-PAM6 | class-3EMat-link | class-PAM6-3EMat |
|---|---|---|---|
| N Válido | 302 | 302 | 302 |
| Média | 0,000 | -0,063 | 0,001 |
| Mediana | -0,027 | -0,132 | 1,031 |
| Desvio padrão | 1,002 | 1,146 | 0,477 |
| Coeficiente de assimetria | -0,126 | 0,673 | 0,477 |
| Erro padrão da Assimetria | 0,140 | 0,140 | 0,140 |
| Amplitude | 7,742 | 7,111 | 6,839 |
| Mínimo | -4,225 | -2,642 | -2,752 |
| Máximo | 3,517 | 4,469 | 4,086 |
| Percentil 25 | -0,593 | -0,840 | -0,696 |
| Percentil 75 | 0,604 | 0,638 | 0,706 |

$0,05$, não se rejeita a hipótese nula sob a qual não existem diferenças estatisticamente significativas entre as classificações na PAM6 e no teste 3EMat.

O coeficiente da correlação de Pearson entre os resultados na PAM6 (percent-

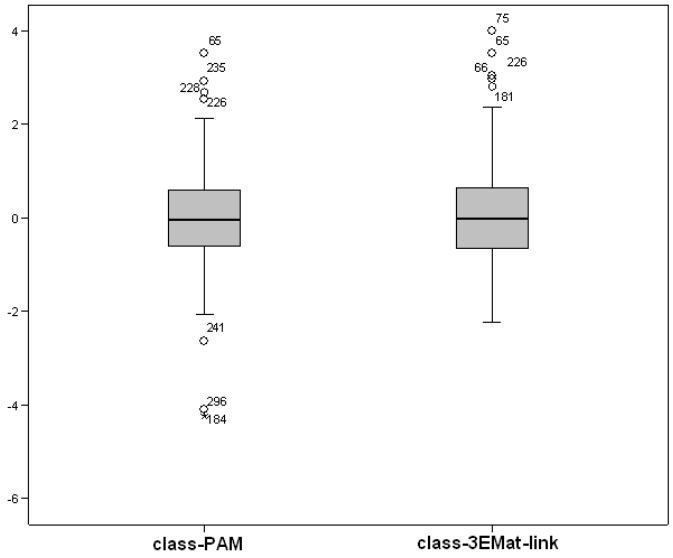



Figura 3.12: Diagramas de extremos e quartis das classificações

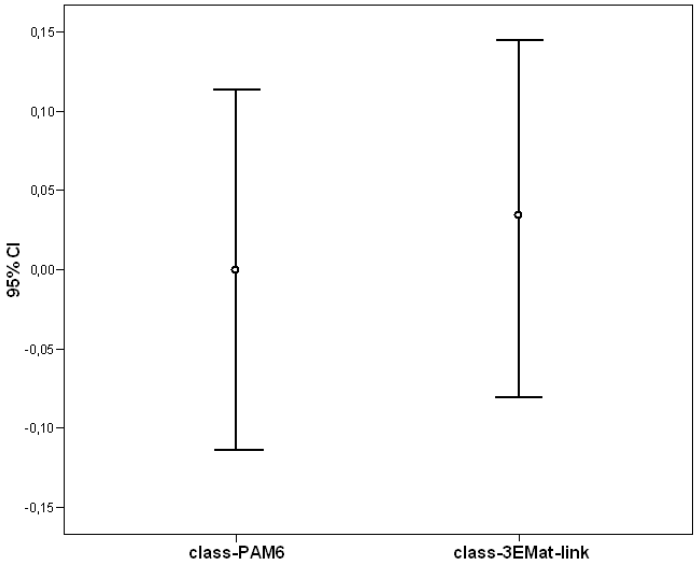



Figura 3.13: Gráficos dos intervalos de confiança para a média das classificações

PAM6) e todas as classificações obtidas anteriormente são apresentados na tabela 3.25.

Verifica-se que é moderada a correlação entre a class-3EMat-link e a class-PAM6, bem como com os resultados na PAM6. É de notar que, a correlação entre class-

Tabela 3.25: Correlação de Pearson entre os resultados e as classificações

| | percent-PAM6 | class-PAM6 | class-3EMat-link | class-PAM6-3EMat |
|---|---|---|---|---|
| percent-PAM6 | 1 | 0,968 | 0,751 | 0,946 |
| class-PAM6 | 0,968 | 1 | 0,743 | 0,955 |
| class-3EMat-link | 0,751 | 0,743 | 1 | 0,901 |
| class-PAM6-3EMat | 0,946 | 0,955 | 0,901 | 1 |

PAM6-3Emat e as restantes classificações é forte, sinalizando as potencialidades da estimação conjunta para estabelecer a métrica, na perspectiva da comparação ao longo do tempo.

Com vista à comparação do desempenho a Matemática na PAM6 e no teste 3EMat, aplicámos o teste de hipóteses para amostras emparelhadas. O valor da estatística de teste é $t = -0,828$ e o respectivo valor de prova é $p = 0,408$, pelo que não se rejeita a hipótese de diferença nula entre as classificações dos alunos em cada um dos instrumentos em análise.

### Discussão

O *linking* entre as escalas obtidas pela aplicação da PAM6 e do teste 3EMat, foi efectuado pelo método linear e pelo método baseado na estimação conjunta dos parâmetros dos itens e do factor latente, assumindo que cada um dos instrumentos (PAM e 3EMat) é um subteste aplicado à mesma amostra.

A correlação obtida entre as quatro escalas estudadas é moderada a forte. A correlação entre a classificação obtida através da estimação conjunta e as restantes classificações é forte, mostrando ser promissor o método aplicado para estabelecer a métrica na perspectiva de comparação dos resultados escolares ao longo do tempo.

A análise dos resultados obtidos permitiu-nos verificar que não existem diferenças estatisticamente significativas entre as classificações obtidas na PAM6 e no teste 3EMat.

Esta aplicação evidencia outra das vantagens dos MRI que é possibilitar a com-

paração entre alunos da mesma população que tenham sido submetidos a instrumentos totalmente diferentes.

# Capítulo 4

# Aplicações - Modelos de Resposta ao Item Multidimensionais

Neste capítulo começamos por usar os modelos de resposta ao item multidimensionais para analisar a dimensionalidade de um teste. Seguidamente, aplicamos o procedimento de estimação proposto no âmbito deste trabalho (Capítulo 2 - secção 2.3), para a estimação dos parâmetros dos modelos multidimensionais, a dados simulados. Na última secção deste capítulo, usamos o procedimento de estimação proposto em dados reais, com vista, a identificar o número de factores latentes que o teste afere e comparamos os resultados obtidos com os do software comercial Testfact (Wilson, Wood e Gibbons [115]). Uma breve descrição de cada uma das aplicações é apresentada seguidamente.

Na secção 4.1, usamos o modelo multidimensional logístico de 3 parâmetros com vista a analisar a dimensionalidade de um teste de Matemática aplicado a alunos do 9º ano, do 3º Ciclo do Ensino Básico. Os dados foram obtidos no âmbito do projecto de investigação 3EM. Apresentamos as técnicas utilizadas para a análise de dimensionalidade de um teste, nomeadamente, os métodos de análise factorial de informação restrita e de informação plena. Os resultados mostram que o teste é unidimensional e que os itens do teste se ajustam melhor ao modelo de resposta ao

item logístico de 3 parâmetros.

Seguidamente, apresentamos os resultados das análises com dados simulados, obtidos pela aplicação do algoritmo MCMC proposto, ao modelo compensatório multidimensional logístico de 2 parâmetros (equação 2.2.1). Para isso, usamos o algoritmo de *Metropolis-Hastings* com amostragem de *Gibbs*. O algoritmo proposto permite a obtenção das estimativas dos parâmetros dos itens e dos factores latentes do modelo. Neste algorimto, a estimação de todos os parâmeros do modelo é feita simultaneamente. Nesta secção, começamos por apresentar os resultados da análise de dados simulados considerando que os dados aferem 2 factores latentes. Seguidamente, mostramos os resultados obtidos no caso em que a dimensão do factor latente é 3. Para ambos os casos, descrevemos a forma como são gerados os dados, utilizamos critérios e estatísticas que permitem comparar os resultados das simulações (critério de informação de Akaike, correlação, erro absoluto médio e erro quadrático médio) e apresentamos as principais conclusões. Os resultados obtidos mostraram que, em ambos os casos, o algoritmo proposto é eficaz em termos computacionais e de estimação dos parâmetros.

Na secção 4.3, temos como propósito analisar o número de dimensões de um instrumento constituído por itens de Matemática e itens que avaliam a Percepção do Autoconceito Infantil - PAI, aplicado a uma amostra de alunos do 1º ano do 1º Ciclo do Ensino Básico da Região da Cova da Beira (projecto 3EM). A estrutura desta secção é a seguinte. Começamos por apresentar a interpretação de cada item a partir de algumas estatísticas da TCT. Seguidamente, para a análise de dimensionalidade do instrumento, utilizamos os métodos de análise factorial de informação restrita e plena. Nesse sentido, efectuamos a inspecção dos valores próprios da matriz de correlações tetracóricas e analisamos as cargas das dimensões rotacionadas. Para a análise dos dados foram utilizados o software Testfact (Wilson, Wood e Gibbons [115]) e o programa em Matlab, que recorre à abordagem bayesiana proposta no âmbito deste trabalho tese. Os resultados obtidos pela aplicação do modelo de resposta

ao item multidimensional mostram que o instrumento afere 3 factores latentes.

## 4.1 Análise da dimensionalidade de um teste

### Dados e Resultados

Os dados utilizados nesta aplicação referem-se a uma amostra aleatória de 277 alunos do 9º ano de escolaridade e foram recolhidos no âmbito do projecto 3EM (Anexo 3 - secção 3.3). O teste foi aplicado no final do ano lectivo 2006/2007.

A interpretação de cada item do teste foi feita, inicialmente, a partir algumas estatísticas da TCT. A tabela 4.1 apresenta o índice de dificuldade e a correlação bisserial de cada item que compõe o teste.

Tabela 4.1: Estatísticas da TCT

| Item | Índice de dificuldade | Correlação Bisserial | Item | Índice de dificuldade | Correlação Bisserial |
|---|---|---|---|---|---|
| 1 | 0,704 | 0,285 | 18 | 0,220 | 0,394 |
| 2 | 0,361 | 0,501 | 19 | 0,217 | 0,455 |
| 3 | 0,473 | 0,566 | 20 | 0,466 | 0,638 |
| 4 | 0,390 | 0,551 | 21 | 0,152 | 0,674 |
| 5 | 0,404 | 0,507 | 22 | 0,058 | 0,676 |
| 6 | 0,329 | 0,366 | 23 | 0,321 | 0,370 |
| 7 | 0,542 | 0,319 | 24 | 0,321 | 0,209 |
| 8 | 0,841 | 0,359 | 25 | 0,635 | 0,517 |
| 9 | 0,462 | 0,407 | 26 | 0,343 | 0,394 |
| 10 | 0,484 | 0,486 | 27 | 0,101 | 0,588 |
| 11 | 0,181 | 0,404 | 28 | 0,682 | 0,349 |
| 12 | 0,383 | 0,355 | 29 | 0,329 | 0,463 |
| 13 | 0,484 | 0,537 | 30 | 0,307 | 0,492 |
| 14 | 0,466 | 0,381 | 31 | 0,466 | 0,475 |
| 15 | 0,390 | 0,591 | 32 | 0,177 | 0,226 |
| 16 | 0,368 | 0,480 | 33 | 0,224 | 0,205 |
| 17 | 0,773 | 0,426 | - | - | - |

Os resultados mostram que o teste apresenta 2 itens considerados fáceis (com

índices de dificuldade acima de 0,75 - itens 8 e 17) e 8 itens considerados difíceis (com índices de dificuldade inferiores a 0,25 - itens 11, 18, 19, 21, 22, 27, 32 e 33). A média do índice de dificuldade é de 0,40. Verifica-se que a correlação bisserial não apresenta valores inferiores a 0,2, tendo os seus valores variado entre 0,205 e 0,676 e a média é de 0,44.

Para mensurar a fiabilidade do teste, foi utilizado o coeficiente de Kuder-Richardson KR20 (Capítulo 1 - secção 1.1). A escala produzida para os 33 itens apresenta consistência interna medida pelo coeficiente KR20 de 0,75, que é considerado um valor adequado.

Para se aplicarem os modelos unidimensionais de resposta ao item, é necessário verificar a predominância de uma dimensão. Como as respostas aos itens estão em formato dicotómico, a investigação da unidimensionalidade do teste é feita a partir das ferramentas disponíveis no software Testfact 2.13 (Wilson, Wood e Gibbons [115]).

O primeiro passo do método de informação restrita consiste na inspecção dos valores próprios da matriz de correlação tetracórica. Na tabela 4.2, são apresentados os primeiros 11 valores próprios e na figura 4.1 a respectiva representação gráfica.

Tabela 4.2: Valores próprios da matriz de correlação tetracórica

| Dimensão | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Valor próprio | 11,26 | 2,64 | 2,07 | 2,04 | 1,85 | 1,72 | 1,66 | 1,42 | 1,27 | 1,20 | 1,04 |

Verifica-se que o primeiro valor próprio é 11,26 e o segundo é 2,64, ou seja, o primeiro é cerca de 4 vezes superior ao segundo. O terceiro valor próprio é de apenas 2,07, sendo pouco inferior ao segundo, e assim sucessivamente. Como o primeiro valor próprio é superior aos demais e a partir da segunda dimensão os valores próprios apresentam valores próximos, podemos inferir que há um valor próprio dominante extraído da matriz de correlação tetracórica.

Com vista a complementar a inspecção dos valores próprios utiliza-se a estatís-

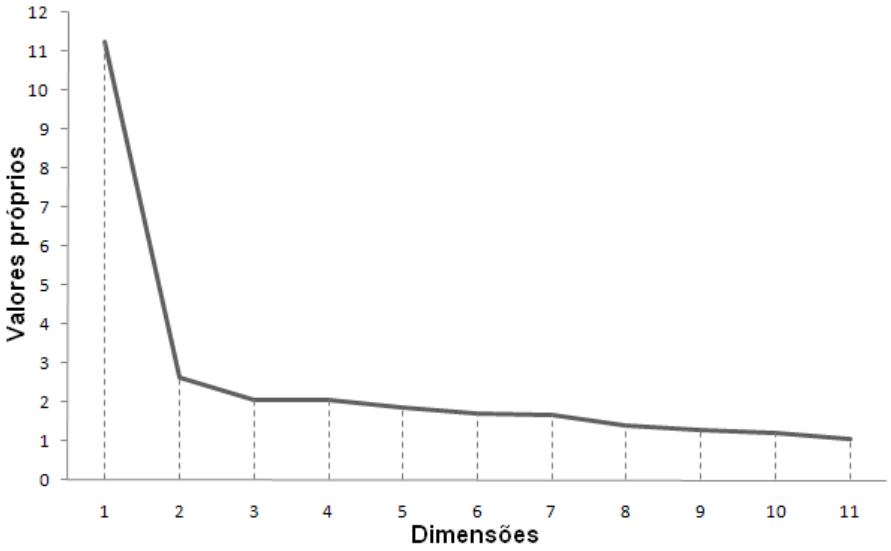



Figura 4.1: Valores próprios da matriz de correlação tetracórica considerando 11 dimensões para o factor latente

tica $G^2$ (Bock, Gibbons e Muraki [17]). Esta estatística permite verificar se existe diferença estatisticamente significativa no aumento da percentagem de variância explicada pelo acréscimo de uma dimensão. Os resultados obtidos apresentam-se na tabela 4.3.

Tabela 4.3: Análise da dimensionalidade

| Número de factores | Variância Associada | | Diferença de $G^2$ | g.l. | Sig. |
|---|---|---|---|---|---|
| | Dimensão 1 | Dimensão 2 | | | |
| 1 | 50,966 | - | - | - | - |
| 2 | 48,543 | 9,524 | 40,44 | 32 | 0,146 |

A percentagem de variância explicada para os 2 factores é, respectivamente, 48,54% e 9,52%. Por esta análise já é possível constatar um ajuste adequado do modelo aos dados apenas com um factor. Acrescenta-se o facto, do acréscimo de explicação com a solução a 2 factores não ser estatísticamente significativa. Contudo, será apresentado o segundo factor com vista a confirmar-se se este é interpretável.

Para o primeiro e o segundo valores próprios é apresentada a matriz de cargas

na tabela 4.4.

Tabela 4.4: Matriz de cargas dos dois primeiros factores baseada no método de informação restrita

| Item | Dimensões | | Item | Dimensões | |
|---|---|---|---|---|---|
| | 1 | 2 | | 1 | 2 |
| 1 | 0,200 | -0,066 | 18 | 0,577 | -0,441 |
| 2 | 0,639 | 0,074 | 19 | 0,678 | -0,421 |
| 3 | 0,615 | 0,330 | 20 | 0,723 | 0,349 |
| 4 | 0,591 | 0,035 | 21 | 0,870 | -0,018 |
| 5 | 0,649 | 0,104 | 22 | 0,726 | -0,103 |
| 6 | 0,517 | 0,098 | 23 | 0,544 | -0,166 |
| 7 | 0,358 | 0,266 | 24 | 0,359 | 0,159 |
| 8 | 0,207 | 0,204 | 25 | 0,569 | 0,110 |
| 9 | 0,592 | 0,461 | 26 | 0,504 | 0,315 |
| 10 | 0,522 | 0,107 | 27 | 0,654 | 0,255 |
| 11 | 0,567 | 0,166 | 28 | 0,399 | -0,307 |
| 12 | 0,502 | -0,435 | 29 | 0,676 | -0,135 |
| 13 | 0,658 | -0,180 | 30 | 0,542 | -0,029 |
| 14 | 0,436 | -0,526 | 31 | 0,559 | -0,081 |
| 15 | 0,671 | 0,069 | 32 | 0,510 | -0,168 |
| 16 | 0,667 | -0,037 | 33 | 0,556 | -0,182 |
| 17 | 0,356 | 0,262 | - | - | - |

Pela análise da tabela, podemos verificar que todos os itens possuem cargas positivas para o primeiro factor, o que significa que à medida que o primeiro factor aumenta, aumenta também a probabilidade de o item ser respondido correctamente. Adicionalmente, verificamos que a quase totalidade dos itens apresenta cargas superiores a 0,30, que é considerado por alguns autores (Johnson e Wichern [67]) um valor mínimo para que se possa considerar o item na interpretação do factor. Já o segundo factor apresenta cargas positivas e negativas sem, aparentemente, nenhum padrão lógico. Em particular, em 5 itens a carga é inferior a -0,3, em 4 itens é superior a 0,3 e nos restantes 24 itens a carga apresenta valores desprezíveis, cujos valores da carga em módulo não são superiores a 0,3. Pode-se concluir que o segundo valor próprio não mede qualquer informação relevante. Assim, pela aplicação do método de in-

formação restrita, pode-se concluir que uma dimensão, aprendizagem/competências em Matemática, é dominante perante as demais.

Como uma forma de se complementar a análise feita anteriormente, utiliza-se a análise factorial com o método de informação plena. A tabela 4.5 apresenta algumas estatísticas obtidas para este método.

Tabela 4.5: Estatísticas da análise baseada no método de informação plena

| | P. discriminação | | Matriz de cargas | | | P. discriminação | | Matriz de cargas | |
|---|---|---|---|---|---|---|---|---|---|
| Item | Dimensões | | Dimensões | | Item | Dimensões | | Dimensões | |
| | $1^a$ | $2^a$ | $1^a$ | $2^a$ | | $1^a$ | $2^a$ | $1^a$ | $2^a$ |
| 1 | 0,220 | 0,066 | 0,207 | 0,083 | 18 | 2,392 | -1,551 | 0,834 | -0,442 |
| 2 | 0,853 | -0,215 | 0,652 | -0,105 | 19 | 4,961 | -3,481 | 0,854 | -0,494 |
| 3 | 1,025 | 0,206 | 0,693 | 0,204 | 20 | 1,594 | -0,231 | 0,848 | -0,048 |
| 4 | 1,231 | 0,080 | 0,768 | 0,118 | 21 | 3,134 | -0,218 | 0,953 | 0,017 |
| 5 | 1,040 | -0,453 | 0,711 | -0,238 | 22 | 1,107 | -0,551 | 0,724 | -0,284 |
| 6 | 1,721 | 1,101 | 0,711 | 0,549 | 23 | 6,794 | -2,885 | 0,943 | -0,306 |
| 7 | 0,404 | 0,315 | 0,333 | 0,311 | 24 | 0,309 | -0,240 | 0,306 | -0,197 |
| 8 | 0,543 | 0,969 | 0,305 | 0,678 | 25 | 0,974 | 0,081 | 0,689 | 0,119 |
| 9 | 1,081 | 0,440 | 0,676 | 0,347 | 26 | 0,599 | -0,170 | 0,519 | -0,099 |
| 10 | 0,961 | 0,094 | 0,683 | 0,128 | 27 | 1,080 | 1,101 | 0,533 | 0,648 |
| 11 | 0,798 | 0,104 | 0,612 | 0,135 | 28 | 0,368 | -0,101 | 0,351 | -0,064 |
| 12 | 1,129 | -0,436 | 0,741 | -0,213 | 29 | 1,747 | 0,630 | 0,799 | 0,370 |
| 13 | 1,016 | -0,295 | 0,713 | -0,141 | 30 | 0,691 | 0,360 | 0,518 | 0,331 |
| 14 | 2,037 | -1,799 | 0,755 | -0,557 | 31 | 0,827 | -0,357 | 0,635 | -0,210 |
| 15 | 1,303 | 0,109 | 0,783 | 0,135 | 32 | 13,019 | -0,030 | 0,993 | 0,085 |
| 16 | 1,183 | -0,409 | 0,758 | -0,190 | 33 | 2,283 | 0,773 | 0,846 | 0,372 |
| 17 | 0,874 | -0,324 | 0,658 | -0,180 | - | - | - | - | - |

A análise da tabela permite constatar que o primeiro factor apresenta todos os valores do parâmetro de discriminação positivos, com mediana em torno de 1,08, enquanto o segundo factor apresenta parâmetro de discriminação com valores positivos e negativos sem, aparentemente, nenhum padrão lógico, tal como foi verificado no método de informação restrita.

Os resultados obtidos pela análise baseada no método de informação restrita e pelo método de informação plena, permitem concluir que o primeiro factor é do-

minante, pelo que se verifica o pressuposto da unidimensionalidade dos modelos de resposta ao item unidimensionais. Verificado este pressuposto, a independência local fica subjacente (Lord [77] e Lord e Novick [78]), pelo que podem-se ajustar os dados um modelo de resposta ao item unidimensional.

Com vista a verificar qual o modelo que melhor se ajustava aos dados, foi realizado o mesmo tipo de análise da dimensionalidade considerando o modelo multidimensional de 2 parâmetros. Para este modelo, a percentagem de variância explicada do primeiro factor foi de 15,73% e do segundo factor foi de 5,35%. Assim, pela comparação dos resultados obtidos com os da tabela 4.3, podemos verificar que diminuiu substancialmente a percentagem de variância explicada por cada factor, comparativamente com o modelo multidimensional de 3 parâmetros. A expressiva perda de informação do modelo multidimensional de 2 parâmetros comparativamente com o de 3 parâmetros é um forte indício de que o parâmetro de probabilidade de acerto ao acaso (Lord [77]) deve ser considerado na análise dos resultados.

Nesse sentido, foi ajustado o modelo logístico de 3 parâmetros aos dados. Para tal, foi utilizado o software Bilog-MG (Zimowski *et al.* [118]) no cálculo dos parâmetros dos itens e do factor latente dos alunos. Na tabela 4.6, são apresentadas as estimativas dos parâmetros dos itens obtidas pela aplicação do modelo.

A análise da tabela permite verificar que os itens com valores mais elevados do parâmetro de discriminação são 19, 21, 23 e 32 (valores superiores a 1,4) enquanto que os itens menos discriminativos são 1, 7, 8, 24, 28 e 30 (valores inferiores a 0,7). O teste é constituído, essencialmente, por itens discriminativos, apresentando discriminação média de 0,984. Relativamente ao parâmetro de dificuldade, constata-se que os itens mais difíceis são 11, 18, 22, 24, 27, 32 e 33 (valores superiores a 2) e que os itens mais fáceis são 8 e 17 (valores inferiores a -0,75). O teste apresenta itens de todos os níveis de dificuldade, embora se constate que um número significativo de itens exige um elevado grau do factor latente para a sua resolução. A dificuldade média do teste é 1,217. No que se refere à probabilidade de acerto ao acaso, os itens

Tabela 4.6: Estimativas dos parâmetros dos itens

| Item | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | Item | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0,370 | -0,572 | 0,285 | 18 | 1,228 | 2,022 | 0,174 |
| 2 | 0,904 | 1,239 | 0,204 | 19 | 1,446 | 1,769 | 0,160 |
| 3 | 1,103 | 0,506 | 0,194 | 20 | 1,383 | 0,386 | 0,157 |
| 4 | 1,206 | 0,880 | 0,195 | 21 | 1,591 | 1,672 | 0,090 |
| 5 | 0,955 | 0,929 | 0,202 | 22 | 1,293 | 2,502 | 0,045 |
| 6 | 1,004 | 1,870 | 0,257 | 23 | 1,541 | 1,800 | 0,270 |
| 7 | 0,506 | 0,946 | 0,310 | 24 | 0,526 | 2,981 | 0,258 |
| 8 | 0,505 | -1,644 | 0,282 | 25 | 0,974 | -0,026 | 0,262 |
| 9 | 0,967 | 1,051 | 0,297 | 26 | 0,702 | 1,610 | 0,206 |
| 10 | 0,988 | 0,874 | 0,292 | 27 | 0,780 | 2,583 | 0,058 |
| 11 | 0,811 | 2,304 | 0,122 | 28 | 0,526 | -0,148 | 0,317 |
| 12 | 0,998 | 1,583 | 0,285 | 29 | 0,899 | 1,608 | 0,220 |
| 13 | 0,984 | 0,622 | 0,233 | 30 | 0,695 | 1,644 | 0,169 |
| 14 | 1,033 | 1,435 | 0,361 | 31 | 0,832 | 1,009 | 0,279 |
| 15 | 1,221 | 0,893 | 0,199 | 32 | 1,404 | 2,478 | 0,161 |
| 16 | 1,204 | 1,305 | 0,250 | 33 | 1,065 | 2,804 | 0,207 |
| 17 | 0,822 | -0,761 | 0,271 | - | - | - | - |

7, 14 e 28 apresentam valores superiores a 0,3 por possuírem alta probabilidade de acerto ao acaso. O teste apresenta um valor médio da probabilidade de acerto ao acaso igual a 0,22. Os itens do teste são questões com quatro opções de resposta, podendo-se afirmar que a probabilidade de um examinando de baixo factor latente acertar o item é aproximadamente igual a 0,25 (1/(número de opções de resposta)).

A distribuição das estimativas do factor latente está representada no gráfico 4.2 e algumas estatísticas descritivas são apresentadas na tabela 4.7.

Tabela 4.7: Estatísticas descritivas da escala do factor latente

| | |
|---|---|
| Mínimo | -1,397 |
| Máximo | 3,172 |
| Média | -0,007 |
| Desvio padrão | 0,856 |
| Coeficiente de assimetria | 0,778 |
| Erro de assimetria | 0,146 |

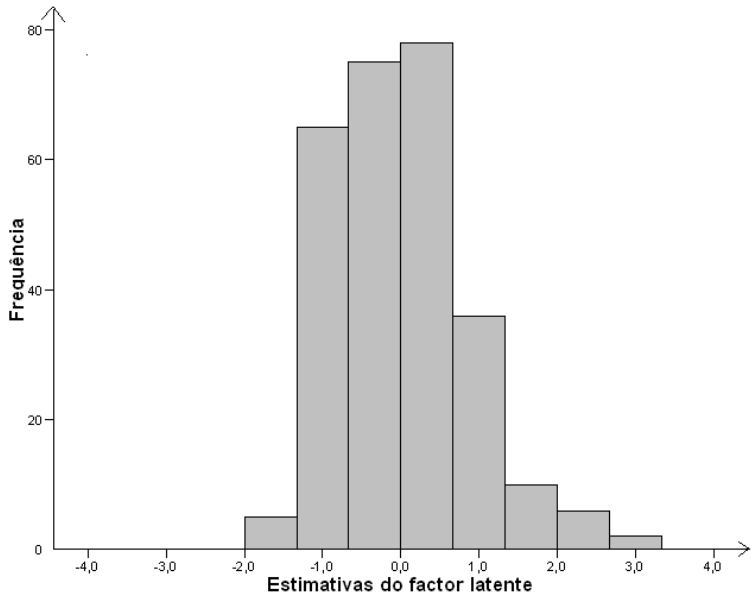



Figura 4.2: Histograma das estimativas do factor latente/competência a Matemática

Verifica-se que esta distribuição é assimétrica positiva (coeficiente de assimetria igual a 0,778 e erro associado igual a 0,146), apesar da distribuição se aproximar de uma normal, com média -0,007 e desvio padrão 0,856.

### Discussão

Nesta aplicação análisámos a dimensionalidade de um teste de Matemática. Apresentámos os resultados obtidos pelos métodos clássicos e pelos métodos baseados em modelos de resposta ao item. Verificámos o pressuposto de unidimensionalidade, recorrendo de forma complementar a dois métodos: método de informação restrita e método de informação plena. Neste contexto usámos modelos de resposta ao item multidimensionais. Ambos os métodos indicaram que o modelo é unidimensional, ou seja, o teste afere aprendizagens a Matemática. Nesse sentido, e verificado este pressuposto, obtivémos a escala do factor latente, competência em Matemática, para o 9º ano pela aplicação do modelo logístico de 3 parâmetros.

A análise dos pressupostos dos modelos unidimensionais de resposta ao item

é essencial em avaliação educacional e decisiva para garantir a qualidade de todo o processo subsequente. Assim, esta aplicação é muito relevante para estudos futuros, uma vez que, a metodologia adoptada poderá ser aplicada a outros testes, nomeadamente, aos aplicados no âmbito do projecto 3EM que aferem aprendizagens a Matemática desde o 1º ao 9º ano de escolaridade do Ensino Básico.

## 4.2 Modelos de resposta ao item multidimensionais com dados simulados

### Dados e Resultados

O algoritmo proposto neste trabalho (capítulo 2, secção 2.3) conjuga a estimação bayesiana com o uso de métodos de simulação de *Markov Chain Monte Carlo* e vai ser testado usando dados simulados para 2 e 3 dimensões do factor latente.

### Caso 1: Duas dimensões para os factores latentes

Foram simuladas as respostas de 2000 examinandos a um instrumento constituído por 40 itens dicotómicos. Para a obtenção das respostas dos examinandos aos itens, geraram-se aleatoriamente os valores verdadeiros de $a$, $d$ e $\theta$, considerando 2 factores latentes. Os valores considerados para os parâmetros foram obtidos através de múltiplas simulações e foram seleccionados de forma que as distribuições se aproximem dos valores das distribuições dos parâmetros do modelo. O parâmetro $d \sim N(\mu_d = 0, c_d^2 = 1, 4^2)$. Para o parâmetro $a$ foram considerados, para os dois factores, os valores de 0 e 1, em 20 itens, 1 e 0 em 10 itens, e, 1 e 1 para os restantes itens. Os factores latentes foram gerados a partir de uma distribuição normal multivariada, com vector média zero, variância 1 e com covariância entre os dois factores 0,3. Foi calculada a probabilidade de resposta correcta de cada examinando a cada item, pela aplicação do modelo multidimensional logístico de 2 parâmetros (equação

2.2.1), e foram geradas as respostas dicotómicas a partir da distribuição binomial. A base de dados correspondente foi inserida no programa em Matlab, a partir do qual foram obtidas as estimativas dos parâmetros dos itens $a$ e $d$ e as estimativas dos factores latentes considerando um número diferente de factores. A distribuições *a priori* utilizadas foram as definidas na subsecção 2.3 do capítulo 2. No algoritmo proposto, para cada iteração, foram obtidas as estimativas dos factores latentes de cada examinando, seguidamente foi estimado o parâmetro de dificuldade de cada item e no final foram obtidas as estimativas dos parâmetros de discriminação de cada item em cada dimensão do factor latente. Para obter as estimativas dos parâmetros do modelo, factores latentes e parâmetros dos itens, foi gerado o parâmetro independentemente para cada examinando ou item, foram calculadas as respectivas probabilidades de aceitação e foi aceite ou rejeitado o novo valor proposto de cada parâmetro considerando o valor da probabilidade obtido anteriormente. Em particular, foram geradas 15000 amostras, descartando as 14000 primeiras para *burn-in*. As estimativas do factor latente e dos parâmetros dos itens foram obtidas através do cálculo da média dos valores de cada parâmetro nas últimas 1000 iterações.

Para a seleção dos modelos em termos do número de dimensões dos factores latentes foi utilizado o critério de informação *Akaike information criterion* (AIC) (Akaike [3]) considerando 1, 2 e 3 factores. O critério AIC é dado por:

$$AIC = -2log(L(U|\Theta, A, d)) + 2K \tag{4.2.1}$$

onde,

$K$ é o número de parâmetros estimados pela aplicação do modelo;

A função de verosimilhança e os restantes parâmetros já foram definidos no capítulo 2.

Os valores obtidos para este critério constam na tabela 4.8.

Com o recurso ao critério AIC, um modelo é mais apropriado quanto menor for essa estatística. Assim, pela análise da tabela 4.8, o número de factores que deve

Tabela 4.8: AIC para dados simulados

| Número de factores | AIC |
|---|---|
| 1 | $1,327 \times 10^5$ |
| 2 | $7,774 \times 10^4$ |
| 3 | $8,103 \times 10^4$ |

ser considerado e que proporciona um melhor ajuste dos dados é dois.

Para dois factores latentes, foram comparados os valores verdadeiros com as estimativas obtidas para os parâmetros do modelo. Para a comparação do parâmetro de discriminação recorremos ao coeficiente de correlação KR20 (Capítulo 1 - secção 1.1) e para os restantes parâmetros procedemos ao cálculo da correlação de Pearson (tabela 4.9). A representação gráfica dos diagramas de dispersão para os valores verdadeiros e as estimativas obtidas é apresentada nas figuras 4.3, 4.4 e 4.5.

Tabela 4.9: Correlação entre os valores verdadeiros e as estimativas obtidas

| | Correlação |
|---|---|
| Corr($a_1$,$\hat{a}_1$) | 0,770 |
| Corr($a_2$,$\hat{a}_2$) | 0,808 |
| Corr($d$,$\hat{d}$) | 0,994 |
| Corr($\theta_1$,$\hat{\theta}_1$) | 0,777 |
| Corr($\theta_2$,$\hat{\theta}_2$) | 0,880 |

A correlação entre os valores verdadeiros dos parâmetros de discriminação e as estimativas obtidas é considerada adequada. No que se refere ao parâmetro de dificuldade dos itens, a correlação é de 0,99. Relativamente à comparação entre valores verdadeiros dos dois factores latentes e as estimativas obtidas, a correlação varia entre 0,78 e 0,88. Assim, verifica-se que, a correlação entre as estimativas dos parâmetros de dificuldade e dos factores latentes com os valores verdadeiros é positiva forte, o que é corroborado pela análise dos diagramas de dispersão.

Seguidamente, são apresentados dois exemplos da comparação das estimativas dos parâmetros dos itens, obtidas nas últimas 1000 iterações do algoritmo proposto,

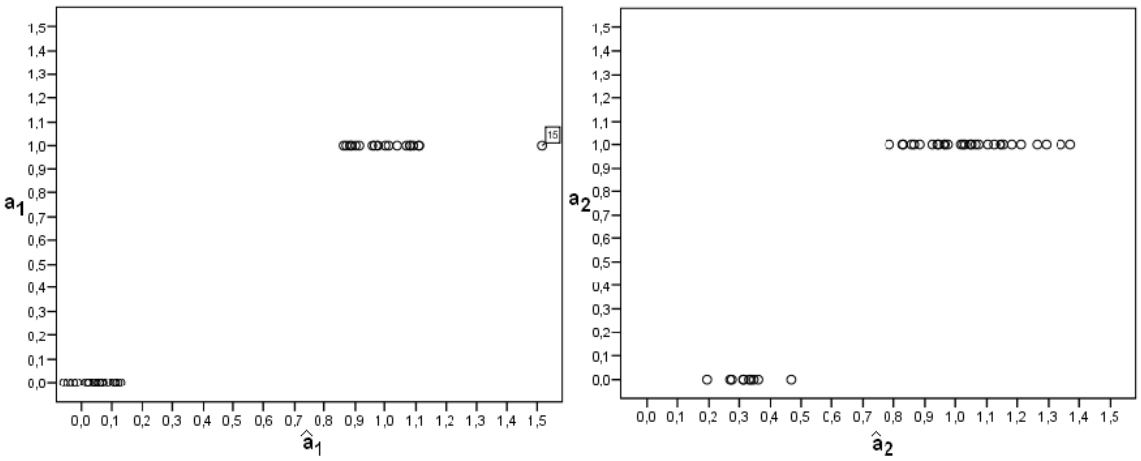



Figura 4.3: Diagramas de dispersão entre os valores verdadeiros e as estimativas obtidas para os parâmetros de discriminação

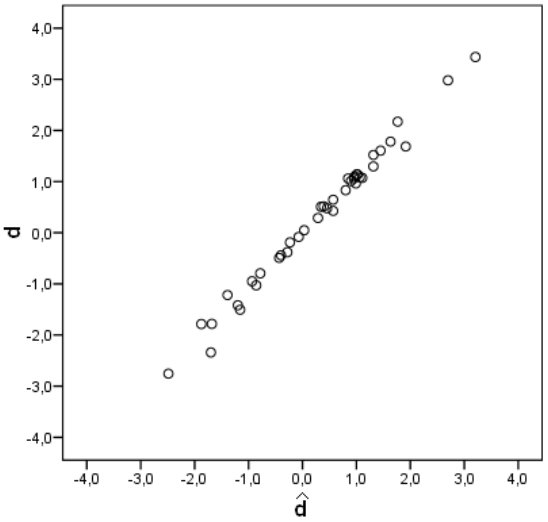



Figura 4.4: Diagramas de dispersão entre os valores verdadeiros e as estimativas obtidas para o parâmetro de dificuldade

com os valores verdadeiros dos parâmetros. Começamos por apresentar um exemplo de um item em que existe um bom ajuste, item 20, e de um item que apresenta um mau ajuste (item 39) em todos os parâmetros dos itens.

Nas figuras 4.6 e 4.7 estão exemplos de representações gráficas das estimativas dos parâmetros do item 20 obtidas pela aplicação do algorimo nas últimas 1000 iterações

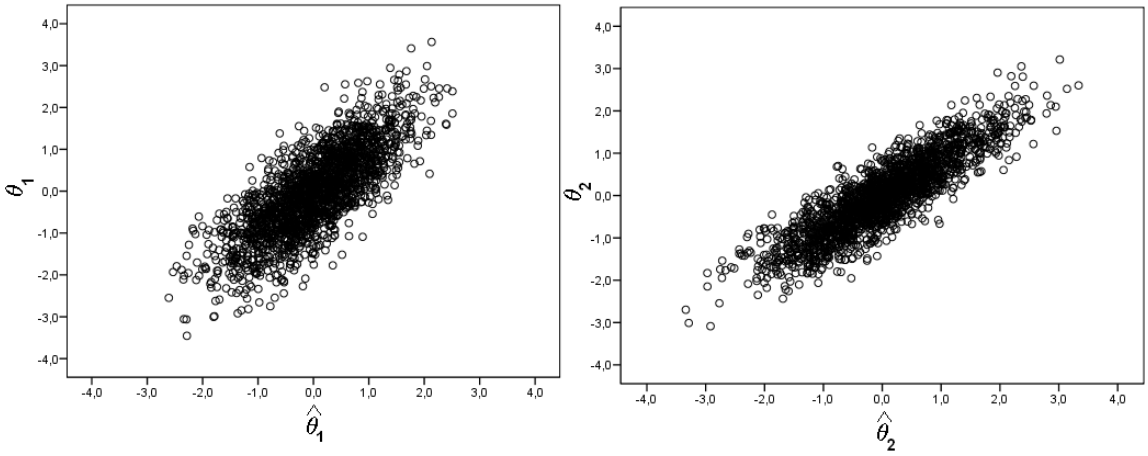



Figura 4.5: Diagramas de dispersão entre os valores verdadeiros e as estimativas obtidas para os factores latentes

e dos valores verdadeiros de cada parâmetro. O valor verdadeiro do parâmetro de discriminação do item 20 para o primeiro factor é 1 e para o segundo factor é 0. Este item é de dificuldade média e o valor verdadeiro do parâmetro de dificuldade é -0,02.

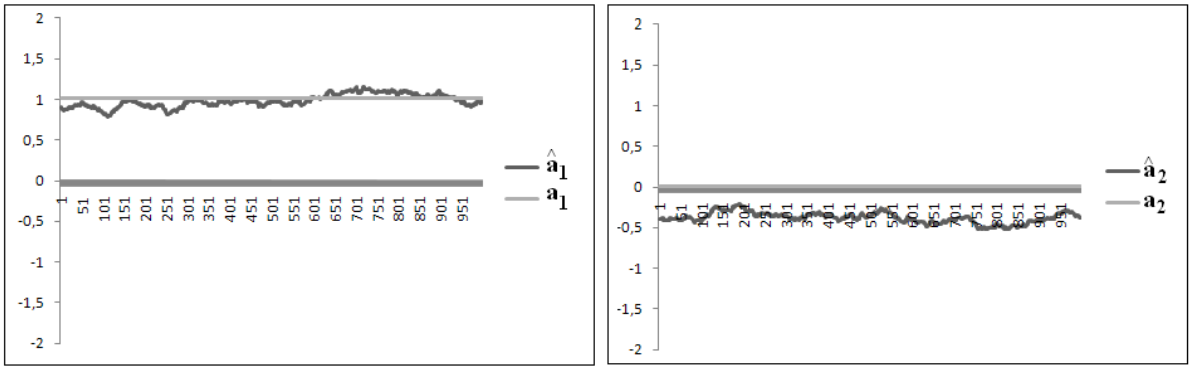



Figura 4.6: Parâmetros de discriminação do item 20 - Comparação das estimativas obtidas nas 1000 iterações e do valor verdadeiro nos 2 factores

As representações gráficas correspondentes ao item 39 são apresentadas nas figuras 4.8 e 4.9. Para este item, o valor verdadeiro do parâmetro de discriminação para o primeiro factor é 1 e para o segundo factor é 0. Este item é fácil e o valor

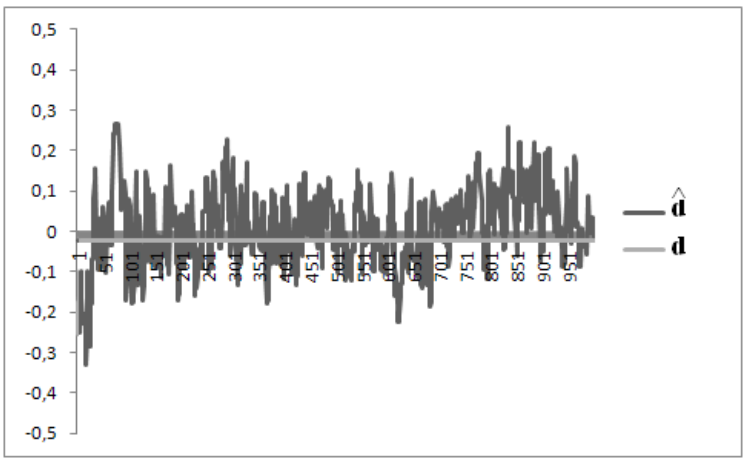



Figura 4.7: Parâmetro de dificuldade do item 20 - Comparação das estimativas obtidas nas 1000 iterações e do valor verdadeiro

verdadeiro do parâmetro de dificuldade é -3,4.

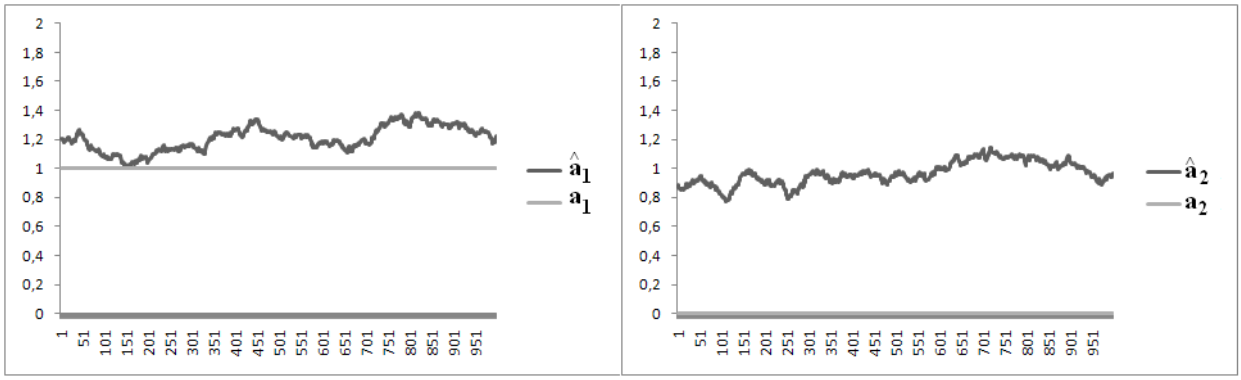



Figura 4.8: Parâmetros de discriminação do item 39 - Comparação das estimativas obtidas nas 1000 iterações e do valor verdadeiro nos 2 factores

A estimativa dos parâmetros de cada item obtém-se pela média das estimativas dos parâmetros nas últimas 1000 iterações do algoritmo. No item 20 os valores obtidos aproximam-se dos valores verdadeiros dos parâmetros do item, o que não acontece no item 39.

As estatísticas Erro Absoluto Médio (EAM) e o Erro Quadrático Médio (EQM) foram calculadas, com vista, a comparar as estimativas obtidas para os parâmetros

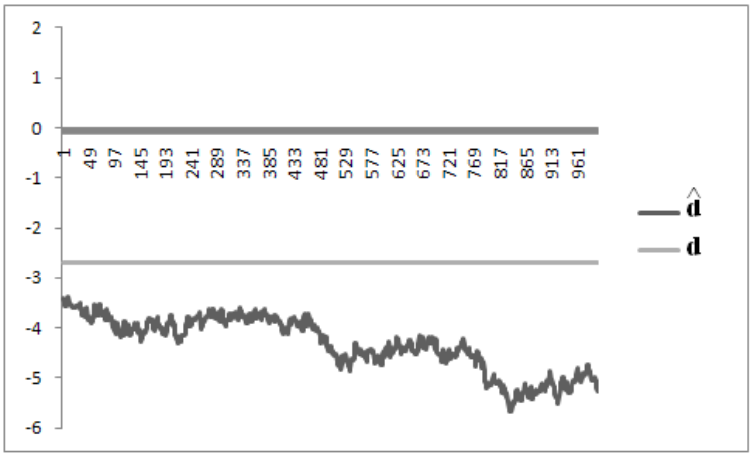



Figura 4.9: Parâmetro de dificuldade do item 39 - Comparação das estimativas obtidas nas 1000 iterações e do valor verdadeiro

dos itens e para os factores latentes, com os valores verdadeiros de cada parâmetro. Os valores obtidos para estas estatísticas são apresentados na tabela 4.10.

Tabela 4.10: Estatísticas EAM e EQM

| | EAM | EQM |
|---|---|---|
| $(a_1,\hat{a}_1)$ | 0,080 | 0,013 |
| $(a_2,\hat{a}_2)$ | 0,177 | 0,046 |
| $(d,\hat{d})$ | 0,136 | 0,035 |
| $(\theta_1,\hat{\theta}_1)$ | 0,509 | 0,410 |
| $(\theta_2,\hat{\theta}_2)$ | 0,375 | 0,223 |

A análise da tabela permite constatar que, para todos os parâmetros do modelo, o EQM apresenta valores inferiores comparativamente com o EAM. Em particular, o EAM varia entre $0,08$ e $0,509$ e o EQM apresenta valores compreendidos entre $0,013$ e $0,41$. Adicionalmente, para ambas as estatísticas, os valores mais elevados são os que correspondem à comparação das estimativas dos dois factores latentes com os valores verdadeiros.

Os tempos de execução do programa em Matlab, num computador Intel core 2, 2 gb de RAM e processador t7200 a 2,00 Ghz, são os seguintes: para estimar os

parâmetros do modelo considerando 1 factor latente o programa demora 32m 58s, para 2 factores latentes o tempo de execução do programa é 33m 31s e para 3 factores latentes o tempo é de 34m 53s. Em geral, quanto mais factores forem considerados maior é o tempo de execução do programa. Aparentemente, o tempo de execução obtido pela aplicação do algoritmo MCMC proposto é inferior ao obtidos por Jiang [66] (Computador com 512 MB RAM e processador 3300 AMD Athlon 64). O algoritmo desenvolvido por este autor para o exemplo mais simples, modelo logístico unidimensional de 3 parâmetros, demorou 37 minutos no caso em que considera a resposta de 2000 examinandos a 30 itens, e um total de 11000 iterações. Assim, para este caso, constata-se que o número de parâmetros estimados e o número de iterações consideradas por Jiang são inferiores ao caso que aqui apresentamos, embora as características dos computadores utilizados sejam diferentes, facto que também influencia a velocidade de execução do programa.

Uma das críticas feitas à aplicação de algoritmos MCMC é o tempo de execução dos programas, que depende, da própria eficiência do programa, do número de examinandos e de itens, da velocidade de convergência e das características do computador utilizado. Nesta aplicação, verifica-se que, também comparativamente com outros softwares como WINBUGS (Bolt e Lall [21]) e SPLUS (Patz e Junker [92] e [93]), o tempo de execução é inferior aos demais, apesar de os computadores usados não apresentarem as mesmas características.

### Caso 2: Três dimensões para os factores latentes

A base de dados simulada é composta pelas respostas de 2500 examinandos a um instrumento constituído por 80 itens dicotómicos. Neste caso, foram considerados 3 factores latentes. Para a obtenção das respostas dos examinandos aos itens geraram-se aleatoriamente os valores verdadeiros de $a$, $d$ e $\theta$. Foram consideradas para o parâmetro $a$, as sequências de valores apresentadas na tabela 4.11.

O parâmetro $d \sim N(\mu_d = 0, c_d^2 = 1, 4^2)$. Os factores latentes foram gerados a

Tabela 4.11: Parâmetros de discriminação verdadeiros

| Sequência | Número de itens |
|---|---|
| 011 | 16 |
| 010 | 24 |
| 111 | 12 |
| 110 | 8 |
| 101 | 12 |
| 100 | 8 |

partir de uma distribuição normal multivariada, com vector média zero, variância 1 e em que a covariância entre os factores é 0,3. A probabilidade de resposta correcta de cada examinando a cada item foi calculada a partir da aplicação da equação 2.2.1, que corresponde ao modelo multidimensional logístico de 2 parâmetros. Foi considerada a distribuição binomial para obter as respostas dicotómicas dos examinandos aos itens. A base de dados foi inserida no programa em Matlab proposto e foram obtidas as estimativas dos parâmetros de discriminação e de dificuldade dos itens e as estimativas dos factores latentes. A descrição do algoritmo utilizado, bem como, das distribuições *a priori* consideradas estão definidas no capítulo 2, secção 2.3. Para o estudo de simulação foram geradas 15000 amostras, descartando as 14000 primeiras para *burn-in*. Foi calculada a média das estimativas de cada parâmetro nas últimas 1000 iterações para obter as estimativas finais.

Para apurar o número de factores que melhor se adequa aos dados foi calculado o critério de informação AIC (tabela 4.12).

Tabela 4.12: AIC para dados simulados

| Número de factores | AIC |
|---|---|
| 1 | $8,351 \times 10^5$ |
| 2 | $1,738 \times 10^5$ |
| 3 | $1,718 \times 10^5$ |
| 4 | $1,746 \times 10^5$ |

A análise da tabela permite concluir que, o critério AIC apresenta valor inferior

para o caso em que o número de factores é 3. Assim, pode-se concluir que devem ser considerados 3 factores latentes para a análise dos dados.

Com o propósito de comparar as estimativas dos parâmetros de discriminação com os valores verdadeiros foi usado o coeficiente de correlação KR20 (Capítulo 1 - secçao 1.1) e para relacionar as estimativas dos parâmetros de dificuldade dos itens e dos factores latentes com os valores verdadeiros foi calculada a correlação de Pearson. Os resultados obtidos são apresentados na tabela 4.13. Adicionalmente, são apresentadas os gráficos de dispersão entre os valores verdadeiros e as estimativas obtidas para os parâmetros do modelo (4.10, 4.11 e 4.12).

Tabela 4.13: Correlação entre os valores verdadeiros e as estimativas obtidas

| | Correlação |
|---|---|
| $\text{Corr}(a_1,\hat{a}_1)$ | 0,601 |
| $\text{Corr}(a_2,\hat{a}_2)$ | 0,723 |
| $\text{Corr}(a_3,\hat{a}_3)$ | 0,607 |
| $\text{Corr}(d,\hat{d})$ | 0,724 |
| $\text{Corr}(\theta_1,\hat{\theta}_1)$ | 0,835 |
| $\text{Corr}(\theta_2,\hat{\theta}_2)$ | 0,568 |
| $\text{Corr}(\theta_3,\hat{\theta}_3)$ | 0,743 |

Em geral, constata-se que os valores obtidos para os parâmetros de discriminação são adequados. A correlação entre os valores verdadeiros e as estimativas obtidas para o parâmetro de dificuldade dos itens é positiva forte. O mesmo se verifica, para a correlação entre as estimativas dos factores latentes e os valores verdadeiros destes parâmetros, com excepção do segundo factor.

As estatísticas erro absoluto médio e erro quadrático médio são apresentadas na tabela 4.14.

Em geral, o EQM apresenta valores inferiores comparativamente com o EAM, com excepção do parâmetro de dificuldade. Os parâmetros de discriminação dos itens apresentam o EAM e o EQM mais baixos e o parâmetro de dificuldade dos itens apresenta o maior erro comparativamente com o respectivo valor verdadeiro.

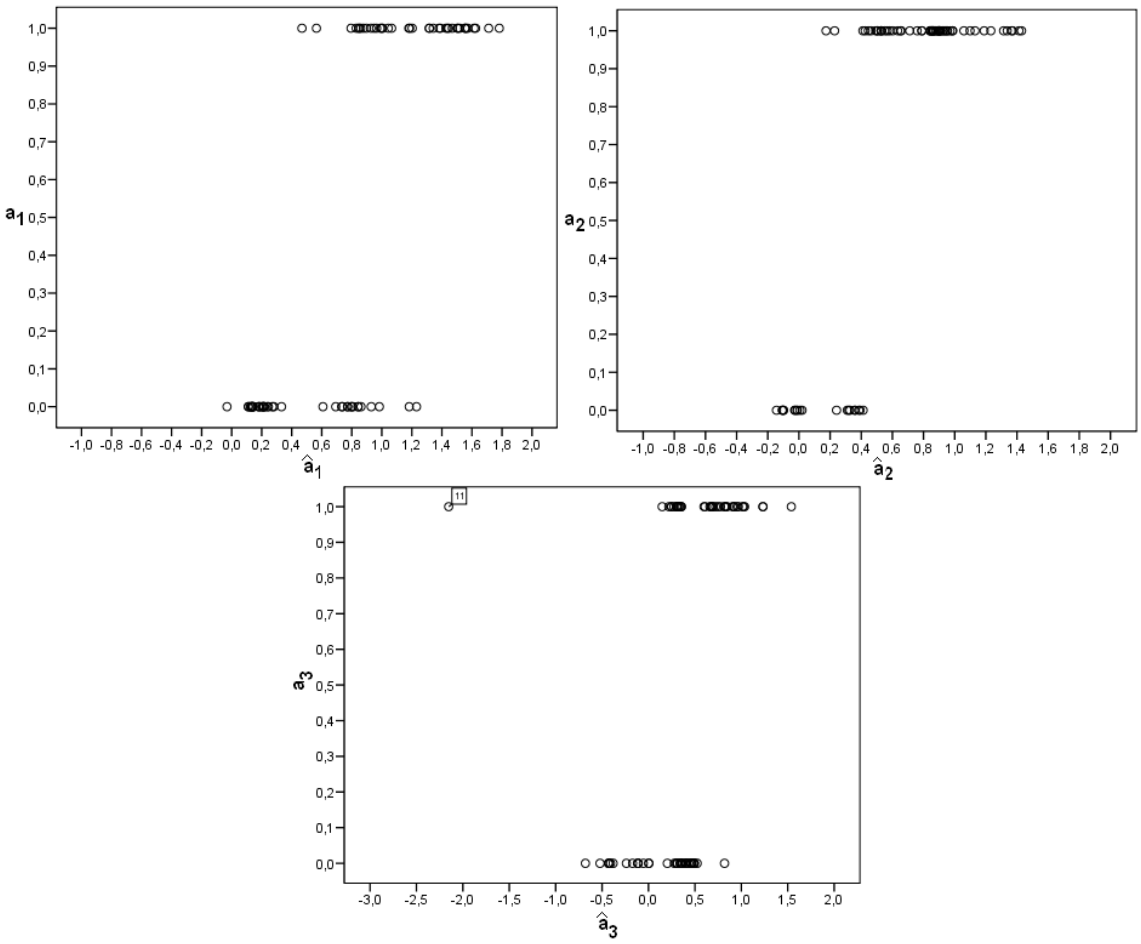



Figura 4.10: Diagramas de dispersão entre os valores verdadeiros e as estimativas obtidas para os parâmetros de discriminação

Tabela 4.14: Estatísticas EAM e EQM

| | EAM | EQM |
|---|---|---|
| $(a_1,\hat{a}_1)$ | 0,247 | 0,097 |
| $(a_2,\hat{a}_2)$ | 0,409 | 0,312 |
| $(a_3,\hat{a}_3)$ | 0,385 | 0,236 |
| $(d,\hat{d})$ | 1,161 | 3,509 |
| $(\theta_1,\hat{\theta}_1)$ | 0,431 | 0,298 |
| $(\theta_2,\hat{\theta}_2)$ | 0,686 | 0,744 |
| $(\theta_3,\hat{\theta}_3)$ | 0,550 | 0,485 |

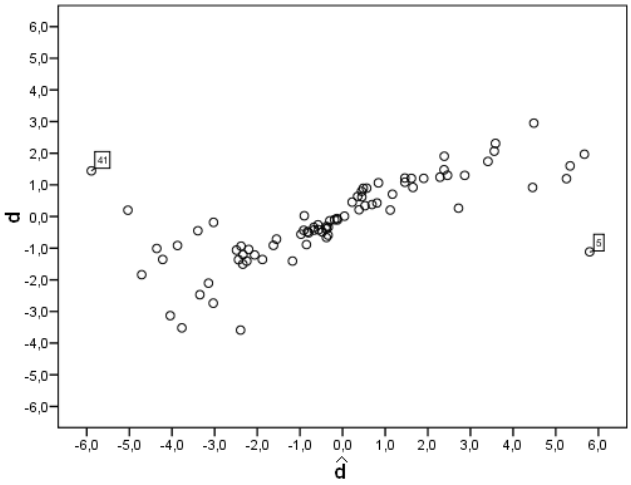



Figura 4.11: Diagramas de dispersão entre os valores verdadeiros e as estimativas obtidas para o parâmetro de dificuldade

Os tempos de execução do programa em Matlab, da base de dados com as respostas de 2500 examinandos a 80 itens, considerando 1, 2, 3 e 4 factores latentes são 1h 54m, 1h 56m, 1h 58m e 1h 59 m, respectivamente. As características do computador utilizado são: Intel core 2, 2 gb de RAM e processador t7200 a 2,00 Ghz.

Tal como seria de esperar, quanto mais factores são considerados maior é o tempo de execução do programa. Considerando diferentes números de factores (entre 1 e 4) verifica-se que a variação dos tempos de execução do programa foi 5 minutos.

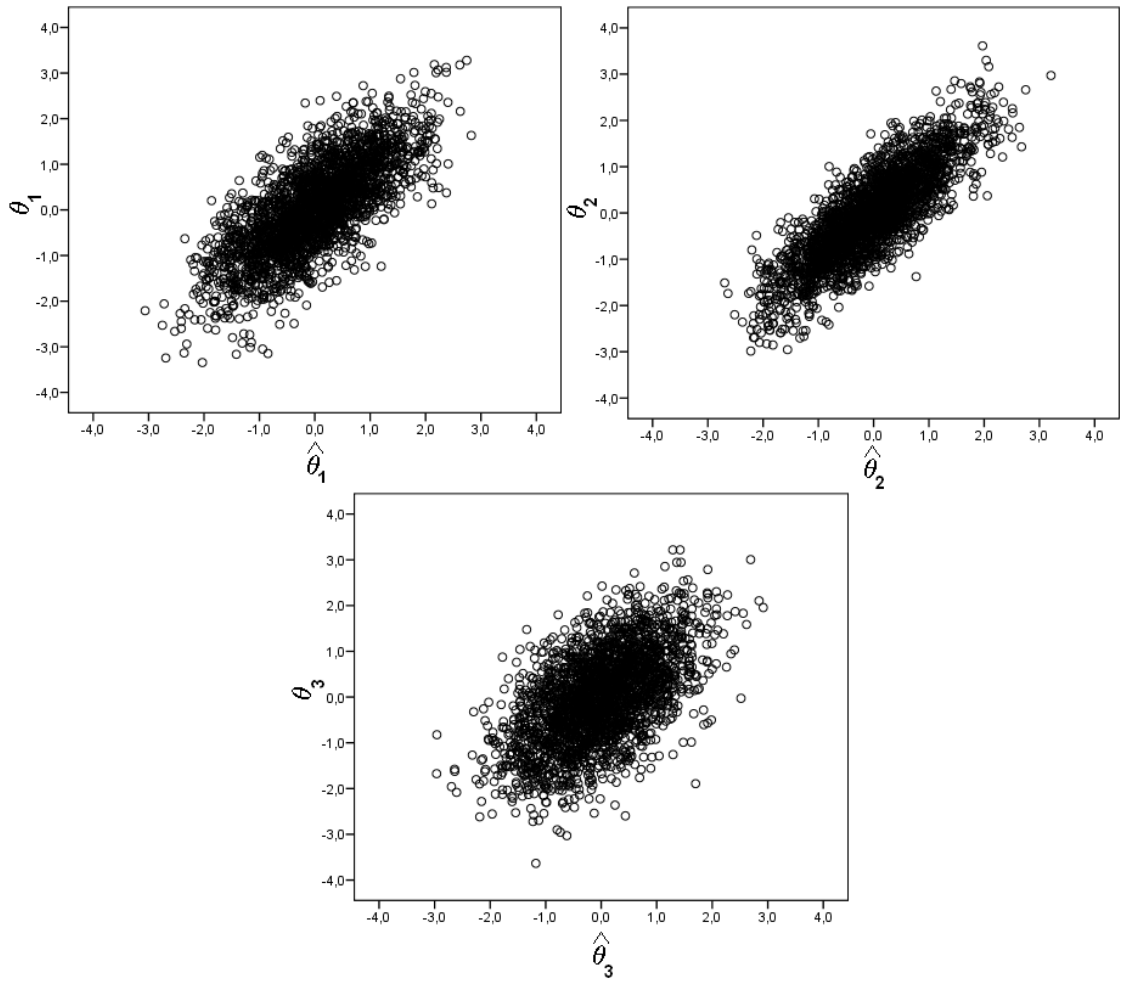



Figura 4.12: Diagramas de dispersão entre os valores verdadeiros e as estimativas obtidas para os factores latentes

### Discussão

Nesta secção, aplicámos a dois conjuntos de dados simulados o modelo logístico multidimensional de 2 parâmetros, com vista, a estudar as suas características e verificar os resultados da abordagem bayesiana proposta, recorrendo a MCMC, na inferência sobre os parâmetros do modelo. Para isso, considerámos que o factor latente aferia duas e três dimensões. Utilizámos o critério de informação AIC com

vista a identificar o número de factores que melhor se ajustava aos dados. Definido o número de factores, para comparar as estimativas dos parâmetros obtidas pela aplicação do modelo, com os valores verdadeiros de cada parâmetro, calculámos: o coeficiente de correlação KR20, a correlação de Pearson, o erro absoluto médio e o erro quadrático médio.

A análise dos resultados de ambas as simulações permitiram-nos concluir que se obtêm boas estimativas para todos os parâmetros do modelo compensatório multidimensional logístico de 2 parâmetros e que a metodologia adoptada para fazer inferência é eficaz para estimar os parâmetros do modelo. Os valores iniciais escolhidos foram obtidos através da realização de simulações. O algoritmo proposto permite estimar simultaneamente os parâmetros associados aos itens e aos factores latentes dos examinandos e pode ser generalizado a outros modelos. Em particular, para a generalização deste modelo ao multidimensional logístico de 3 parâmetros, bastará apenas considerar o parâmetro de probabilidade de acerto ao acaso como parâmetro do item. Este parâmetro é geral para cada item, e como tal, a sua estimação pode ser efectuada como se tratasse de um modelo unidimensional. Outra das vantagens da utilização desta abordagem é a possibilidade da sua aplicação a dados reais. Este estudo vai ser feito na secção seguinte. Para a utilização da abordagem proposta podemos, ainda, apontar as seguintes vantagens: a possibilidade da generalização do algoritmo a outros modelos multidimensionais, por exemplo, os modelos não compensatórios e a aplicação desta abordagem a dados que aferem um maior número de factores latentes.

Ainda no âmbito desta secção, apresentámos os tempos de execução do programa proposto. Em geral, os tempos que obtivémos são inferiores aos apresentados por Jiang [66], que também utiliza a abordagem bayesiana, para a estimação de parâmetros de modelos multidimensionais. Este autor utiliza uma função que estima a matriz de covariância, o que parece tornar o algoritmo mais demorado. Contudo, as características dos computadores utilizados não são iguais.

## 4.3 Modelos de resposta ao item multidimensionais com dados reais

### Dados e Resultados

Os dados foram obtidos no âmbito de um projecto 3EM e referem-se a uma amostra de 309 alunos do 1º ano do Ensino Básico (Anexo 3 - secção 3.3). O instrumento utilizado é composto por 30 itens, que aferem competências a Matemática, e 33 itens que avaliam a Percepção do Autoconceito Infantil - PAI. O teste de Matemática e o PAI foram aplicados no início do ano lectivo de 2005/2006.

A interpretação de cada item do instrumento foi feita, inicialmente, a partir das estatísticas clássicas: índice de dificuldade e correlação bisserial. Na tabela 4.15 são apresentadas essas estatísticas da TCT para cada item do instrumento. Os itens de Matemática são designados por C1 até C30 e os itens do PAI são denominados de Q1 até Q33.

A análise da tabela permite verificar que o instrumento apresenta 33 itens considerados fáceis (com índices de dificuldade acima de 0,75), 22 itens de dificuldade média e 8 itens considerados difíceis (com índices de dificuldade inferiores a 0,25). A média do índice de dificuldade é de 0,657. No que se refere à correlação bisserial, normalmente são aceites valores para esta correlação superiores a 0,2. Pela análise da tabela, podemos constatar que, em geral, os resultados obtidos para a correlação bisserial são superiores a 0,2, com média 0,415 e mediana 0,451.

A escala produzida para os 63 itens apresenta consistência interna medida pelo coeficiente KR20 de 0,795, que é considerado um valor adequado.

Com vista à análise do número de dimensões associadas aos itens do instrumento, recorremos às ferramentas disponíveis no software Testfact(Wilson, Wood e Gibbons [115]). Nesse sentido, começámos por realizar a inspecção dos valores próprios da matriz de correlações tetracóricas. Na tabela 4.16 são apresentados os primeiros 10

Tabela 4.15: Estatísticas da TCT

| Item | Índice de dificuldade | Correlação Bisserial | Item | Índice de dificuldade | Correlação Bisserial |
|---|---|---|---|---|---|
| C1 | 0,343 | 0,536 | Q4 | 0,37 | 0,828 |
| C2 | 0,77 | 0,489 | Q5 | 0,096 | 0,838 |
| C3 | 0,472 | 0,586 | Q6 | 0,509 | 0,828 |
| C4 | 0,809 | 0,52 | Q7 | 0,049 | 0,511 |
| C5 | 0,738 | 0,348 | Q8 | 0,43 | 0,9 |
| C6 | 0,453 | 0,512 | Q9 | 0,523 | 0,939 |
| C7 | 0,084 | 0,677 | Q10 | 0,663 | 0,961 |
| C8 | 0,126 | 0,085 | Q11 | 0,367 | 0,896 |
| C9 | 0,589 | 0,538 | Q12 | 0,254 | 0,964 |
| C10 | 0,353 | 0,52 | Q13 | 0,53 | 0,948 |
| C11 | 0,133 | 0,566 | Q14 | 0,534 | 0,932 |
| C12 | 0,961 | 0,451 | Q15 | 0,351 | 0,929 |
| C13 | 0,482 | 0,439 | Q16 | 0,522 | 0,906 |
| C14 | 0,307 | 0,497 | Q17 | 0,349 | 0,754 |
| C15 | 0,702 | 0,341 | Q18 | 0,505 | 0,893 |
| C16 | 0,029 | 0,096 | Q19 | 0,382 | 0,835 |
| C17 | 0,372 | 0,49 | Q20 | 0,501 | 0,926 |
| C18 | 0,728 | 0,18 | Q21 | 0,172 | 0,877 |
| C19 | 0,356 | 0,411 | Q22 | 0,452 | 0,913 |
| C20 | 0,159 | 0,409 | Q23 | 0,385 | 0,757 |
| C21 | 0,259 | 0,506 | Q24 | 0,578 | 0,922 |
| C22 | 0,252 | 0,339 | Q25 | 0,513 | 0,974 |
| C23 | 0,285 | 0,297 | Q26 | 0,346 | 0,709 |
| C24 | 0,864 | 0,342 | Q27 | 0,392 | 0,945 |
| C25 | 0,625 | 0,572 | Q28 | 0,467 | 0,848 |
| C26 | 0,081 | 0,401 | Q29 | 0,526 | 0,896 |
| C27 | 0,621 | 0,553 | Q30 | 0,55 | 0,929 |
| C28 | 0,107 | 0,674 | Q31 | 0,52 | 0,964 |
| C29 | 0,113 | 0,255 | Q32 | 0,227 | 0,945 |
| C30 | 0,738 | 0,513 | Q33 | 0,437 | 0,871 |
| Q1 | 0,586 | 0,25 | - | - | - |
| Q2 | 0,625 | 0,231 | - | - | - |
| Q3 | 0,916 | 0,192 | - | - | - |

valores próprios.

Tabela 4.16: Valores próprios da matriz de correlação tetracórica

| Dimensão | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Valor próprio | 9,623 | 6,404 | 4,187 | 2,944 | 2,585 | 2,268 | 2,134 | 2,025 | 1,948 | 1,816 |

Pela análise da tabela, podemos verificar que para os 3 primeiros factores os valores próprios são superiores aos demais. Adicionalmente, a partir da quarta dimensão, os valores próprios da matriz de correlação tetracórica apresentam valores próximos. Nesse sentido, é possível constatar a existência de 3 factores dominantes. A percentagem de variância explicada para os 3 primeiros factores é, respectivamente, 15,61%, 10,30% e 7,04%.

A "matriz de cargas" obtida para uma extracção, com rotação pelo método *Promax*, para os três primeiros factores é a que se apresenta na tabela 4.17. O método de rotação usado deve-se ao facto de se supor existir uma correlação significativa entre as possíveis componentes extraídas. Note-se que, Dim. representa Dimensão e que apenas são apresentados os valores de carga que, em valor absoluto, são superiores ou iguais a 0,3, que é considerado um valor mínimo para que se possa considerar o item na interpretação do factor.

A análise destes resultados permite verificar que existem valores de carga, em valor absoluto, superiores ou iguais a 0,3 para os 3 factores. Assim, podemos concluir que existem três factores dominantes, apesar da existência do 3º factor poder ser duvidosa. O primeiro factor (ou dimensão) pode ser interpretado como sendo uma medida da percepção do autoconceito infantil. Este factor está fortemente associado aos itens do PAI (Q1 até Q33), com excepção dos itens Q2, Q7, Q15 e Q17. O segundo e terceiro factores referem-se, essencialmente, a itens que aferem competências a Matemática. No caso do segundo factor, este corresponde aos itens de Matemática (C1 até C30), com excepção dos itens C7, C11, C14, C18, C20, C21, C26, C28 e C29. A análise dos descritores (Ferrão *et al.* [47]) correspondentes aos

Tabela 4.17: Cargas das dimensões rotacionadas

| Item | Dim. 1 | Dim. 2 | Dim. 3 | Item | Dim. 1 | Dim. 2 | Dim. 3 |
|---|---|---|---|---|---|---|---|
| C1 | - | 0,424 | 0,449 | Q4 | -0,336 | - | - |
| C2 | - | 0,523 | - | Q5 | -0,3 | - | - |
| C3 | - | 0,687 | - | Q6 | -0,633 | - | - |
| C4 | - | 0,628 | - | Q7 | - | - | - |
| C5 | - | 0,349 | 0,349 | Q8 | -0,413 | - | - |
| C6 | - | 0,75 | - | Q9 | -0,744 | - | - |
| C7 | - | - | 0,626 | Q10 | -0,582 | - | - |
| C8 | - | - | - | Q11 | -0,529 | - | - |
| C9 | - | 0,606 | - | Q12 | -0,475 | - | - |
| C10 | - | 0,372 | - | Q13 | -0,819 | - | - |
| C11 | - | - | 0,699 | Q14 | -0,425 | - | - |
| C12 | - | 0,576 | - | Q15 | - | - | - |
| C13 | - | 0,515 | - | Q16 | -0,739 | - | - |
| C14 | - | - | 0,705 | Q17 | - | - | - |
| C15 | - | 0,434 | - | Q18 | -0,436 | - | 0,469 |
| C16 | - | -0,509 | 0,744 | Q19 | -0,424 | - | - |
| C17 | - | 0,659 | - | Q20 | -0,582 | - | - |
| C18 | - | - | 0,658 | Q21 | -0,399 | - | - |
| C19 | - | 0,522 | - | Q22 | -0,595 | - | - |
| C20 | - | - | 0,695 | Q23 | -0,332 | - | - |
| C21 | - | - | 0,771 | Q24 | -0,808 | - | - |
| C22 | - | 0,315 | - | Q25 | -0,495 | - | - |
| C23 | - | 0,257 | - | Q26 | -0,3 | - | - |
| C24 | - | 0,386 | - | Q27 | -0,443 | - | - |
| C25 | - | 0,481 | - | Q28 | -0,481 | - | - |
| C26 | - | - | 0,616 | Q29 | -0,573 | - | - |
| C27 | - | 0,766 | - | Q30 | -0,665 | - | - |
| C28 | - | - | 0,734 | Q31 | -0,915 | - | |
| C29 | - | - | - | Q32 | -0,562 | - | - |
| C30 | - | 0,652 | - | Q33 | -0,445 | - | - |
| Q1 | - | - | - | - | - | - | - |
| Q2 | - | - | - | - | - | - | - |
| Q3 | -0,299 | - | - | - | - | - | - |

itens que contribuem para o segundo factor, permite constatar que estes itens se referem a descritores contidos nos temas: Contagem e Números; Forma e Espaço e Grandezas e Medidas, abordados/assimilados no nível do Ensino Pré-Escolar e a alguns descritores iniciais dos três temas previstos para o nível de ensino 1º ano do 1º Ciclo do Ensino Básico. Já o terceiro factor refere-se, na sua maioria, também a itens de Matemática, com excepção dos itens Q13, Q17 e Q18 que pertencem ao PAI. Os itens que contribuem para este factor, em geral, correspondem a conteúdos curriculares que se enquadram nos temas Números e Operações e Forma e Espaço e que são leccionados no 1º ano do Ensino Básico. Como já foi referido anteriormente, este instrumento foi aplicado no início do ano lectivo e talvez por este motivo, aquando da aplicação do instrumento, estes conteúdos curriculares não tinham ainda sido leccionados. Relativamente aos itens Q13, 17, e Q18, são itens que se referem a: como o aluno se sente/afectividade, confiança em si mesmo na realização das tarefas/segurança e sentimento de posse de objectos, respectivamente. Neste caso, não foi possível estabelecer qualquer relação entre estes itens e os de Matemática. A análise da tabela 4.17 permite, ainda, constatar que os itens Q13 e Q18, do PAI, contribuem simultaneamente para os primeiro e terceiro factores e que os itens de Matemática C1, C5 e C8, apresentam cargas superiores a 0,3, tanto para o segundo como para o terceiro factor.

A tabela 4.18 apresenta os parâmetros de discriminação dos itens associados a cada dimensão.

A análise da tabela permite constatar que, em geral, os parâmetros de discriminação dos itens apresentam, em valor absoluto, valores mais elevados no factor a que pertencem. Adicionalmente, considerando o valor absoluto dos parâmetros de discriminação, para os primeiro, segundo e terceiro factores a média é 0,473, 0,382 e 0,281, respectivamente.

Com o propósito de efectuar a análise dos resultados obtidos, utilizámos a abordagem bayesiana proposta pela aplicação do algoritmo desenvolvido em Matlab e

Tabela 4.18: Estimativas dos parâmetros de discriminação para cada dimensão

| Item | Dim. 1 | Dim. 2 | Dim. 3 | Item | Dim. 1 | Dim. 2 | Dim. 3 |
|---|---|---|---|---|---|---|---|
| C1 | 0,539 | -0,699 | -0,157 | Q4 | 0,447 | 0,149 | -0,167 |
| C2 | 0,410 | -0,513 | 0,183 | Q5 | 0,122 | 0,256 | 0,171 |
| C3 | 0,605 | -0,508 | 0,531 | Q6 | 0,682 | 0,451 | 0,092 |
| C4 | 0,448 | -0,499 | 0,415 | Q7 | -0,021 | 0,110 | -0,061 |
| C5 | 0,316 | -0,539 | -0,103 | Q8 | 0,525 | 0,184 | -0,073 |
| C6 | 0,450 | -0,726 | 0,599 | Q9 | 0,795 | 0,682 | 0,114 |
| C7 | 0,754 | -0,503 | -0,466 | Q10 | 0,861 | 0,252 | 0,119 |
| C8 | 0,017 | 0,073 | -0,324 | Q11 | 0,535 | 0,427 | -0,191 |
| C9 | 0,523 | -0,578 | 0,277 | Q12 | 0,342 | 0,359 | 0,139 |
| C10 | 0,433 | -0,245 | 0,124 | Q13 | 0,808 | 0,926 | 0,387 |
| C11 | 0,622 | -0,538 | -0,622 | Q14 | 0,626 | 0,160 | -0,135 |
| C12 | 0,455 | -0,578 | 0,231 | Q15 | 0,375 | 0,053 | -0,125 |
| C13 | 0,343 | -0,317 | 0,358 | Q16 | 0,849 | 0,700 | -0,043 |
| C14 | 0,593 | -0,468 | -0,653 | Q17 | 0,328 | -0,095 | -0,132 |
| C15 | 0,243 | -0,277 | 0,282 | Q18 | 0,860 | 0,173 | -0,443 |
| C16 | -0,002 | -0,092 | -1,249 | Q19 | 0,435 | 0,209 | 0,125 |
| C17 | 0,464 | -0,458 | 0,531 | Q20 | 0,621 | 0,412 | -0,023 |
| C18 | 0,287 | -0,254 | -0,689 | Q21 | 0,238 | 0,380 | -0,083 |
| C19 | 0,324 | -0,276 | 0,425 | Q22 | 0,581 | 0,424 | 0,108 |
| C20 | 0,374 | -0,420 | -0,685 | Q23 | 0,395 | 0,086 | 0,194 |
| C21 | 0,600 | -0,638 | -0,794 | Q24 | 0,977 | 0,783 | 0,339 |
| C22 | 0,226 | -0,264 | 0,099 | Q25 | 0,610 | 0,315 | -0,195 |
| C23 | 0,196 | -0,167 | 0,112 | Q26 | 0,351 | 0,085 | 0,126 |
| C24 | 0,247 | -0,227 | 0,241 | Q27 | 0,471 | 0,228 | 0,068 |
| C25 | 0,599 | -0,390 | 0,144 | Q28 | 0,611 | 0,200 | 0,085 |
| C26 | 0,383 | -0,177 | -0,603 | Q29 | 0,717 | 0,362 | -0,091 |
| C27 | 0,516 | -0,806 | 0,581 | Q30 | 0,690 | 0,540 | 0,034 |
| C28 | 0,794 | -0,460 | -0,719 | Q31 | 1,137 | 1,371 | 0,532 |
| C29 | 0,129 | -0,168 | 0,026 | Q32 | 0,292 | 0,589 | 0,098 |
| C30 | 0,547 | -0,483 | 0,463 | Q33 | 0,484 | 0,219 | 0,081 |
| Q1 | 0,248 | 0,214 | 0,203 | - | - | - | - |
| Q2 | 0,146 | 0,046 | 0,118 | - | - | - | - |
| Q3 | 0,200 | 0,260 | -0,102 | - | - | - | - |

calculámos o critério de informação AIC para comparar os resultados, considerando 1, 2, 3 e 4 factores. Os valores obtidos para este critério são apresentados na tabela 4.19. Para a obtenção destes resultados, foram geradas 15000 amostras, descar-

tando as 14000 primeiras para *burn-in*. As estimativas dos factores latentes e dos parâmetros dos itens foram obtidas através do cálculo da média dos valores de cada parâmetro nas últimas 1000 iterações.

Tabela 4.19: AIC

| Número de factores | AIC |
|---|---|
| 1 | $1,612 \times 10^4$ |
| 2 | $1,564 \times 10^4$ |
| 3 | $1,539 \times 10^4$ |
| 4 | $1,552 \times 10^4$ |

A análise da tabela anterior, permite constatar que o menor valor para o AIC se obtém para 3 factores, o que vem corroborar pela análise feita anterioremente recorrendo ao software Testfact (Wilson, Wood e Gibbons [115]).

### Discussão

Nesta secção analisámos o número de dimensões de um instrumento, aplicado a uma amostra de alunos do 1º ano, que é composto por itens de Matemática e itens que avaliam a Percepção do Autoconceito Infantil - PAI, num total de 63 itens. A amostra de 309 alunos corresponde a alunos que frequentaram o 1º Ciclo do Ensino Básico da Região da Cova da Beira, em Portugal.

Efectuámos a interpretação inicial de cada item que compõe o instrumento a partir de estatísticas da TCT. O instrumento apresentava 33 itens fáceis, 22 itens de dificuldade média e 8 itens difíceis. Em geral, os itens do teste apresentaram correlação bisserial superior a 0,3.

O coeficiente de Kuder-Richardson foi de 0,795 que é considerado um valor adequado para mensurar a fiabilidade do teste.

Para a análise de dimensionalidade do instrumento, utilizámos, inicialmente o método de análise factorial de informação restrita. Para isso, foram obtidos os valores próprios da matriz de correlações tetracóricas, utilizando o software Testfact

(Wilson, Wood e Gibbons [115]). A análise dos valores próprios permitiu verificar a existência de 3 factores dominantes. De forma complementar,utilizámos o método de análise factorial de informação plena. A análise da "matriz de cargas"obtida para uma extracção, com rotação pelo método *Promax*, para os três primeiros factores, permitiu concluir que existem 3 factores latentes dominantes, apesar da interpretação do 3º factor ser duvidosa. O 1º factor pôde ser interpretado como uma medida da percepção do autoconceito infantil. O 2º factor, em geral, corresponde a itens de Matemática que se referem a descritores iniciais para o 1º ano do Ensino Básico e que já foram adquiridos pelos alunos no Ensino Pré-Escolar. O 3º factor refere-se, essencialmente, a itens de Matemática que aferem conteúdos curriculares e que são leccionados apenas no 1º ano. É de notar que, como o instrumento foi aplicado no início do ano lectivo, possivelmente estes conteúdos ainda não tinham sido leccionados.

Adicionalmente, com o propósito de verificar os resultados obtidos foi realizada a análise dos resultados utilizando a abordagem bayesiana proposta, que recorre a MCMC. Os valores do critério de informação AIC vieram corroborar que o modelo se ajusta melhor aos dados, considerando 3 factores.

A utilização da abordagem bayesiana, utilizando *Metropolis-Hastings* com amostragem de *Gibbs* em dados reais é inovadora e os resultados obtidos vieram confirmar que esta abordagem produz boas estimativas para os parâmetros do modelo.

# Conclusões e trabalhos futuros

Os modelos de resposta ao item são utilizados em todo o mundo em diversas áreas de conhecimento. Em Portugal, foram usados os modelos de resposta ao item na dissertação de mestrado (Costa [24]). Nesta dissertação, foram explorados os modelos de resposta ao item unidimensionais para dados dicotómicos. Este trabalho vem no seguimento dessa dissertação e com vista a contribuir para colmatar as dificuldades em termos de aplicabilidade de modelos de resposta ao item, tanto do ponto de vista teórico, devido a problemas de difícil solução no campo da estimação de parâmetros, como do ponto de vista computacional. Nesse sentido, esta tese reflecte o trabalho cujos objectivos foram: a exploração de modelos de resposta ao item politómicos unidimensionais; a utilização de modelos para grupos múltiplos; a aplicação de procedimentos estatísticos de equalização e *linking*; o alargarmento os modelos unidimensionais logísticos de 1, 2 e 3 parâmetros a modelos multidimensionais e a exploração de procedimentos de simulação MCMC, para a optimização dos procedimentos de estimação. Assim, propomos um procedimento de estimação, que conjuga a abordagem bayesiana com MCMC, para obter as estimativas dos parâmetros dos itens e dos factores latentes do modelo multidimensional logístico de 2 parâmetros.

Esta tese estruturou-se da seguinte forma. Começámos por apresentar os modelos de resposta ao item unidimensionais considerando dados dicotómicos e politómicos. Descrevemos os modelos para grupos múltiplos. Explorámos os conceitos das funções de informação do item e do teste e abordámos os procedimentos de equa-

lização e *linking*. Seguidamente, apresentámos a parte teórica relativa aos modelos de resposta ao item multidimensionais, nomeadamente, a revisão da literatura e a especificação formal dos principais modelos multidimensionais de resposta ao item. No que se refere aos procedimentos de estimação, expusémos os que mais se utilizam e as suas limitações, descrevemos os procedimentos de simulação de MCMC em modelos de resposta ao item e propusémos um procedimento de estimação que usa MCMC para a estimação dos parâmetros de modelos multidimensionais. No penúltimo capítulo deste trabalho, efectuámos aplicações dos modelos de resposta ao item unidimensionais. No final deste trabalho explorámos os modelos de resposta ao item multidimensionais.

Nas aplicações dos modelos de resposta ao item unidimensionais, começámos por estudar as propriedades psicométricas de um teste de escolha múltipla, que foi utilizado para aferir competências em Estatística, no âmbito do Mestrado Integrado em Engenharia e Gestão Industrial da Universidade do Minho. Para a análise dos dados, utilizámos as abordagens baseadas na TCT e em MRI. As estatísticas da TCT usadas foram: índice de dificuldade, índice de discriminação e correlação ponto-bisserial. O modelo de resposta ao item unidimensional dicotómico aplicado aos dados foi o logístico de 2 parâmetros. Apresentámos as estimativas do parâmetro de discriminação, as estimativas do parâmetro de dificuldade e a função de informação do teste. A análise dos itens que compunham o teste, considerando ambas as abordagens, indicaram que o teste era constituído por itens, essencialmente, discriminativos e muito discriminativos e a existência de itens de todos os níveis de dificuldade. Identificámos os melhores e os piores itens do instrumento. Ainda, no âmbito desta aplicação, comparámos as classificações dos alunos de anos lectivos subsequentes (2006/2007 e 2007/2008). Os resultados indicaram uma melhoria na média das classificações obtidas pela aplicação do teste em 2007/2008. Esta aplicação permitiu-nos mostrar as vantagens da análise baseada em MRI, nomeadamente, no que diz respeito à criação de bancos de itens e à modelação de dados com vista à obtenção da métrica

que permite a comparação efectiva dos resultados atingidos pelos alunos em anos lectivos diferentes.

Seguidamente, modelaram-se os dados da Prova de Aferição de Matemática do 4º ano de escolaridade, do 1º Ciclo do Ensino Básico, aplicada no ano lectivo 2006/2007, utilizando o modelo de resposta ao item unidimensional de Crédito Parcial Generalizado, para dados politómicos. Para realizar a análise dos itens que compunham a prova, apresentámos as estimativas do parâmetro de discriminação, as estimativas do parâmetro de dificuldade, as estimativas dos parâmetros de intersecção das categorias adjacentes. Recorremos à função de informação do teste para averiguar a capacidade informativa do teste relativamente ao factor latente (competências em Matemática no 4º ano) e quantificámos o erro da medida. Para validar a metodologia de análise, avaliámos a adequação do modelo teórico aos dados, baseada num teste do qui-quadrado. O ajuste do modelo aos dados, na generalidade, mostrou-se adequado. Os resultados obtidos permitiram constatar que: esta prova era constituída, essencialmente, por itens de discriminação média e baixa, de nível de dificuldade médio e baixo e os itens 4 e 14 devem ser revistos ou retirados (itens com baixo poder discriminativo e muito fáceis). A análise da função de informação do teste e do erro padrão da medida permitiram concluir que esta prova continha grande poder informativo relativamente ao factor latente de alunos com nível de desempenho baixo e médio e indicou elevada imprecisão dos resultados obtidos no que se refere à aferição da aprendizagem de alunos com nível de desempenho no extremo superior da escala. Os resultados apurados ressaltaram a importância da utilização de técnicas baseadas em MRI, para a análise das propriedades dos instrumentos e dos itens, em avaliação educacional, assim como o seu potencial na construção e validação de instrumentos.

A aplicação referente ao modelo de resposta ao item para grupos múltiplos teve como propósito utilizar um modelo de resposta ao item multidimensional, para analisar a dimensionalidade de testes aplicados para aferir competências/aprendizagens

a Matemática. Os dados foram obtidos no âmbito do projecto 3EM. A estrutura dos dados exigiu que se utilizasse a modelação baseada em grupos múltiplos, com vista à criação de uma escala única de desempenho a Matemática. O procedimento de equalização via itens comuns permitiu obter, conjuntamente, as estimativas dos parâmetros dos itens e do factor latente e a obtenção de escalas comuns para o factor latente, em todos os anos lectivos e aplicações. O facto de, se obter uma estimativa única dos parâmetros do modelo diminui os erros em termos de equalização comparativamente com a estimação dos parâmetros aplicação a aplicação ou ano a ano.

Na aplicação de modelos de respostas ao item usando procedimentos de equalização, o objectivo foi o de ajustar uma escala vertical única padronizada do desempenho a Matemática no Ensino Básico. Os modelos foram aplicados aos dados longitudinais, recolhidos entre 2004 e 2007, no âmbito do projecto 3EM. Os testes 3EMat, foram aplicados a uma amostra aleatória dos alunos dos 1º e 2º anos de escolaridade do Ensino Básico. As escalas de desempenho, para aferir as competências desenvolvidas em Matemática, nos diferentes níveis do Ensino Básico, foram feitas com base na aplicação do modelo de resposta ao item logístico de 2 parâmetros. O procedimento de estimação de MVM foi utilizado para estimar o factor latente, desempenho em Matemática. Aplicámos o método de equalização via itens comuns utilizando o procedimento Média-Desvio e o procedimento Média-Média. Em ambos os procedimentos, verificámos que houve uma melhoria das classificações do 1º ano para o 2º ano de escolaridade. O erro padrão de medida obtido para as classificações nos itens âncora foi menor no procedimento Média-Média. O trabalho empírico desenvolvido corroborou a evidência registada na literatura de que o procedimento Média-Média produz estimativas dos parâmetros mais estáveis.

Para finalizar as aplicações de modelos de resposta ao item unidimensionais, efectuámos o *linking*/ligação entre as escalas construídas a partir dos testes 3EMat e provas de aferição, considerando a disciplina de Matemática do 6º ano de escola-

ridade (PAM6) no ano lectivo 2006/2007. Para isso, utilizámos o método linear e a estimação conjunta, assumindo que cada um dos instrumentos (PAM6 e 3Emat) eram subtestes aplicados à mesma amostra. Efectuámos uma comparação dos resultados das distribuições marginais das classificações e o estudo da distribuição conjunta, no que se refere ao *matching* de casos válidos, e as escalas PAM6 e 3EMat foram comparadas em termos dos resultados obtidos pelos alunos em ambos os instrumentos. Constatámos, ainda, que não existem diferenças estatisticamente significativas entre as classificações obtidas na PAM6 e no teste 3EMat. As correlações entre as quatro escalas estudadas foram de moderada a forte. O método aplicado mostrou ser promissor para estabelecer a métrica, na perspectiva de comparação dos resultados escolares ao longo do tempo, e evidencia a possibilidade de comparação entre alunos da mesma população que tenham sido submetidos a instrumentos totalmente diferentes.

No que se refere a modelos de resposta ao item multidimensionais, usámos esta classe de modelos para analisar a dimensionalidade de um teste de Matemática aplicado a alunos do 9º ano do 3º Ciclo do Ensino Básico. Para isso, utilizámos os métodos de análise factorial de informação restrita e de informação plena. Os dados foram obtidos no âmbito de um projecto de investigação 3EM. O instrumento utilizado refere-se a um teste de múltipla escolha, constituído por 33 itens, que afere competências a Matemática. Os resultados mostraram que o teste é unidimensional e que os itens do teste se ajustam melhor ao modelo de resposta ao item logístico de 3 parâmetros. A análise dos pressupostos dos modelos unidimensionais de resposta ao item é essencial para a avaliação educacional e decisiva para garantir a qualidade de todo o processo subsequente. Nesse sentido, esta aplicação mostra-se muito relevante para estudos futuros, uma vez que a metodologia adoptada poderá ser aplicada a outros testes.

Seguidamente, apresentámos os resultados das análises com dados simulados, obtidos pela aplicação do algoritmo MCMC proposto ao modelo compensatório mul-

tidimensional logístico de 2 parâmetros (equação 2.4.1). Nesse sentido, começámos por apresentar os resultados considerando que os dados aferem 2 factores latentes. Posteriormente, mostrámos os resultados obtidos no caso em que a dimensão do factor latente é 3. Para ambos os casos, descrevemos a forma como são gerados os dados, foram utilizados critérios e estatísticas que permitiram comparar os resultados das simulações e apresentaram-se as principais conclusões. A análise dos resultados de ambas as simulações permitiram concluir, que se obtêm boas estimativas para todos os parâmetros do modelo compensatório multidimensional logístico de 2 parâmetros, e que a metodologia adoptada para fazer inferência é inovadora na estimação dos parâmetros do modelo. O algoritmo proposto permite estimar simultaneamente os parâmetros associados aos itens e aos factores latentes dos examinandos e pode ser generalizado a qualquer número de factores latentes, bem como a outros modelos. Os tempos que se obtiveram são inferiores aos apresentados por outros autores que também utilizam abordagens bayesianas, apesar das características dos computadores utilizados serem diferentes.

Para finalizar as aplicações, analisámos o número de dimensões de um instrumento aplicado a uma amostra de alunos do 1º ano de escolaridade do Ensino Básico, que é composto por itens de Matemática e itens que avaliam a Percepção do Autoconceito Infantil - PAI, num total de 63 itens. A amostra de 309 alunos era composta por alunos que frequentavam o 1º Ciclo do Ensino Básico da Região da Cova da Beira. A interpretação inicial de cada item que compunha o instrumento foi feita a partir de algumas estatísticas da TCT. O instrumento apresentou 33 itens fáceis, 22 itens de dificuldade média e 8 itens difíceis. Em geral, os itens do teste apresentaram correlação bisserial superior a 0,3. O coeficiente de Kuder-Richardson foi de 0,795 que é considerado um valor adequado para mensurar a fiabilidade do teste. Para a análise da dimensionalidade do instrumento, começámos por utilizar o método de análise factorial de informação restrita. Para isso, foram obtidos os valores próprios da matriz de correlações tetracóricas, utilizando o software Testfact (Wilson, Wood

e Gibbons [115]). A análise dos valores próprios permitiu verificar a existência de 3 factores dominantes. De forma complementar, usámos o método de análise factorial de informação plena. A análise da "matriz de cargas" obtida para uma extracção, com rotação pelo método *Promax*, para os três primeiros factores permitiu concluir que existem 3 factores latentes dominantes, apesar da interpretação do 3º factor ser duvidosa. O 1º factor pôde ser interpretado como uma medida da percepção do autoconceito infantil. O 2º factor, em geral, correspondeu a itens de Matemática que se referiam a descritores iniciais para o 1º ano do Ensino Básico e que já foram adquiridos pelos alunos no Ensino Pré-Escolar. O 3º factor referia-se, essencialmente, a itens de Matemática que aferiam conteúdos curriculares que são leccionados apenas no 1º ano. Realçámos que, como o instrumento foi aplicado no início do ano lectivo, possivelmente estes conteúdos ainda não tinham sido leccionados. Adicionalmente, verificámos os resultados obtidos, utilizando a abordagem bayesiana proposta, que recorre a MCMC. Os valores do critério de informação AIC vieram corroborar que o modelo se ajusta melhor aos dados, considerando 3 factores. A utilização da abordagem bayesiana, utilizando *Metropolis-Hastings* com amostragem *Gibbs* em dados reais é inovadora e os resultados obtidos vieram confirmar que a utilização desta abordagem é eficaz na obtenção dos parâmetros do modelo.

Como limitações do presente trabalho, no algoritmo MCMC proposto, apontam-se os seguintes aspectos:

- Poder-se-iam ter considerado outros valores para o parâmetro de discriminação e não apenas valores inteiros;
- Considerar factores não correlacionados no algoritmo;
- Na utilização de dados simulados, considerar um número maior de dimensões para o factor latente.

Surgiram da reflexão sobre a bibliografia consultada, mas sobretudo no decorrer deste trabalho a indicação de alguns trabalhos futuros:

- A extensão dos modelos multidimensionais (compensatórios e não-compensatórios) a partir de uma modelagem bayesiana que incorpore a detecção simultânea de cargas que sejam estatisticamente diferentes de zero. Essa estrutura adicional, que pode ser obtida a partir de modelos de misturas, como os usados em Soares *et al.* [107], no contexto do Funcionamento Diferencial do Item;

- Modelagem de estruturas de rotação directamente na especificação do modelo e análise das suas implicações sobre as conclusões obtidas pelo modelo.

# Bibliografia


[1] Abrahamowicz, M. e Ramsay, J.O. (1992). Multicategorical spline model for item response theory. *Psychometrika*, 57, 5-27.

[2] Adams, R.J.; Wilson, M. e Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-24.

[3] Akaike, H. (1981). Likelihood of o model and information criteria. *Journal of Econometrics*, 16, 3-14.

[4] Albert, J.H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, 17, 251-269.

[5] Andrade, D.F.; Tavares, H. R. e Valle, R.C. (2000). *Teoria da Resposta ao Item: conceitos e aplicações*. 14º Sinape - Caxambu: Associação Brasileira de Estatística.

[6] Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.

[7] Angoff, W.H. (1971). Scales, norms and equivalent scores. Em *Educational Measurement* - 2nd ed (Thorndike, R.L., Ed.), 508-600. Washington, DC: American Council on Education.

[8] Baker, F.B. (1992). *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker.

[9] Baker, F.B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling approach. *Applied Psychological Measurement*, 22, 153-169.

[10] Baker, F.B. e Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, 28, 147-162.

[11] Baker, F.B. e Kim, S.H. (2004). *Item Response Theory - Parameter Estimation Techniques* - 2nd ed. New York: Marcel Dekker Inc.

[12] Béguin, A.A. e Glas, C.A.W. (2001). MCMC Estimation and Model-fit analysis of Multidimensional IRT Models. *Psychometrika*, 66, 541-562.

[13] Bejar, I. I. (1977). An application of the continuous response level model to personality measurement. *Applied Psychological Measurement*, 1, 509-521.

[14] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Em *Statistical Theories of Mental Test Scores* (Lord, F. e Novick M., Eds.). New York: Addison-Wesley.

[15] Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.

[16] Bock, R.D. e Aitkin M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. *Psychometrika*, 46, n. 4, 443-459.

[17] Bock, R.D.; Gibbons, R.D. e Muraki, E. (1988). Full-Information Factor Analysis. *Applied Psychological Measurement*, 12, 261-280.

[18] Bock, R.D. e Lieberman, M. (1970). Fitting a response model for n dichotomously scored itens. *Psychometrika*, 35, 179-197.

[19] Bock, R.D. e Schilling, S.G. (2003). IRT based item factor analysis. Em *IRT from SSI: Bilog-mg, Multilog, Parscale, Testfact* (Toit, M. du, Ed.), 584-591. Lincolnwood, Illinois: Scientific Software International.

[20] Bock, R.D. e Zimowski, M.F. (1997). Multiple Group IRT. Em *Handbook of Modern Item Response Theory* (van der Linden, W.J. e Hambleton, R.K., Eds.), 433-448. New York: Spring-Verlag.

[21] Bolt, D.M. e Lall, V.F. (2003). Estimation of Compensatory and Noncompensatory Multidimensional Item Response Models using Markov Chain Monte Carlo. *Applied Psychological Measurement*, 27, n. 6, 395-414.

[22] Boring, E.G. (1945). The use of operational definitions in science. *Psychological Review*, 52, 243-245.

[23] Chen W-H.; Revicki, D.A.; Lai, J-S.; Cook, K.F. e Amtmann, D. (2009). Linking Pain Items from Two Studies Onto a Common Scale Using Item Response Theory. *Journal of Pain and Symptom Management*, 38, n. 4, 615-628.

[24] Costa, P. (2006). *Modelos de Resposta ao Item.* Dissertação de Mestrado. Covilhã: Universidade da Beira Interior.

[25] Costa, P. e Ferrão, M.E. (2005). Modelo de Resposta ao Item na Estimação da Qualidade da Infra-estrutura das Escolas. *Actas do XII Congresso Anual da SPE*, 195-206.

[26] Costa, P.; Ferrão, M.E.; Fernandes, N. e Soares, T. (2009). Uma aplicação da análise factorial na detecção das dimensões cognitivas em testes de avaliação em larga escala em Portugal. Em *XLI Simpósio Brasileiro de Pesquisa Operacional - Pesquisa Operacional na Gestão do Conhecimento*, Porto Seguro, Anais XLI SBPO, 1, 391-402. Rio de Janeiro: SOBRAPO.

[27] Costa, P.; Fletcher, P. e Ferrão, M.E. (2007). Modelo de Resposta ao Item de 2 parâmetros: estimação via MML e EM. Em *Estatística: Ciência Jubilar, Actas do XIV Congresso Anual da Sociedade Portuguesa de Estatística* (Ferrão, M.E., Nunes, C. e Braumann, C., Eds), 305-312. Lisboa: Edições SPE.

[28] Costa, P.; Oliveira, P. e Ferrão, M.E. (2008). Equalização de escalas com o modelo de resposta ao item de dois parâmetros. Em *Estatística - da Teoria à Pratica, Actas do XV Congresso Anual da Sociedade Portuguesa de Estatística* (Hill, M.; Ferreira, M.; Dias, J.; Salgueiro, M.; Carvalho, H.; Vicente, P. e Braumann, C., Eds.), 155-166. Lisboa: Edições SPE.

[29] Costa, P.; Oliveira, P. e Ferrão, M. E. (2009). Statistical Issues on Multiple Choice Tests in Engineering Assessment. Em *Proceedings of the 37th Sefi conference* (Bogaard, M., Graaff, E. e Saunders-Smith, G., Eds), Rotterdam: Delft University of Technology.

[30] Costa, P.; Oliveira, P. e Ferrão, M.E. (2009). *Statistical Issues on Multiple Choice Tests in Engineering Assessment: A Classical Test Theory Approach.* Working Paper.

[31] Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

[32] von Davier, A.A.; Holland, P.W. e Thayer, D.T. (2004). *The Kernel Method of Test Equating* - Statistics for Social and Public Policy. New-York: Springer.

[33] Decreto-Lei nº 6/2001. D.R. I Série-A 15 (2001-01-18) 258-262.

[34] Despacho nº 2351/2007. D.R. II Série - Nº 32 (2007-02-14) 3979.

[35] Dempster, A.P.; Laird, N.M. e Rubin, D.B. (1977). Maximum Likelihood from incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39, série B, 1-38.

[36] D'Hainaut, L. (1992). *Conceitos e Métodos da Estatística - Volume II: Duas ou três variáveis segundo duas ou três dimensões*. Lisboa: Fundação Calouste Gulbenkian.

[37] Divgi, D.R. (1979). Calculation of the tetrachoric correlation coefficient. *Psycometrika*, 44, n. 2, 169-172.

[38] Dorans, N.J. (1990). Equating methods and sampling designs. *Applied Measurement in Education*, 3, 3-17.

[39] Dorans, N.J. (2004). Profiles in Research: Ledyard R. Tucker. *Journal of Educational and Behavioral Statistics*, 29, n. 1, Spring, 145-151.

[40] Dorans, N.J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research*, 16, Suppl. 1, 85-94.

[41] Dorans N.J. e Holland, P.W. (2000). Population invariance and the equatability of tests: basic theory and the linear case. *Journal of Educational Measurement*, 37, Issue 4, 281-306.

[42] Dunn, G. (1989). *Design and Analysis of Reliability Studies: the statistical evaluation of measurement errors*. London: Edward Arnold.

[43] ENQA (2005). *Standards and Guidelines for Quality Assurance in the European Higher Education Area.* Helsinki: European Association for Quality Assurance in Higher Education.

[44] Ferrão, M.E.; Costa P.; Navio, V.M. e Dias, V.M. (2006). Medição da competência dos alunos do ensino básico em Matemática: 3EMAT, uma proposta. *Actas da XI Conferência Internacional Avaliação Psicológica: Formas e Contextos*, 905-915. Universidade do Minho.

[45] Ferrão, M.E.; Costa, P. e Oliveira, P.N. (2009). *Item Response Model Applied to Developing a Common Metric: statistics assessment in Engineering*. Working Paper.

[46] Ferrão, M.E.; Costa, P. e Gama, J. (2010). Distribution-free Item Response Model based on Marginal Maximum Likelihood Estimation. *Advances and Applications in Statistics*, 18, n. 2, 109-126.

[47] Ferrão, M.E.; Loureiro, M.J.; Navio, V.M. e Coelho, I. (2009). *Aferição das Aprendizagens a Matemática no Ensino Básico: a proposta 3EMAT*. Covilhã: Universidade da Beira Interior.

[48] Feuer, M.J.; Holland, P.W.; Green, B.F., Bertenthal, M.W. e Hemphill, F.C. (1999). *Uncommon Measures: Equivalence and Linkage Among Educational Tests.* Washington DC: National Academy Press.

[49] Flanagan, J.C. (1951). Units, scores, and norms. Em *Educational Measurement* (Lindquist, E.F., Ed.), 695-763. Washington, DC: American Council on Education.

[50] Gabinete de Avaliação Educacional. *Provas de Aferição.* [Consult. 2008-09-01]. Disponível em http://www.gave.min-edu.pt/np3/7.html.

[51] Gelfand, A.E. e Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85, 398-409.

[52] Geman, S. e Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

[53] Ghadan, K.E. (2005). The effects of sample size on equating of test items. *Education*, 126, Fall, 165-180.

[54] Guilera, G. e Gómez, J. (2008). Item response theory test equating in Health Sciences Education. *Advances in Health Sciences Education*, 13, n. 1, 3-10.

[55] Guilford, J.P. e Fruchter, B. (1978). *Fundamental statistics in psychology and education*, 6th ed, New York: McGraw-Hill.

[56] Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japonese Psychological Research*, 22, 144 -149.

[57] Haley, D.C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error* (Technical Report N$^o$ 15). Stanford: Stanford University Applied Mathematics and Statistics Laboratory, CA.

[58] Hambleton, R. K. e Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston, MA: Kluwer Nijhoff Publishing.

[59] Hambleton, R.K.; Swaminathan, H. e Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. North Caroline: Sage Publications.

[60] Hand, D.J. (2004). *Measurement, Theory and Practice*. London: Arnold.

[61] Harris, D.J. e Crouse, J.D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, 6, 195-240.

[62] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.

[63] Holland, P.W.; Dorans, N.J. e Petersen, N.S. (2007). Equating Test Scores. *Handbook of Statistics*, 26, 169-203.

[64] Holland, P.W. e Rubin, D.B. (1982). *Test Equating*. New-York: Academic Press.

[65] Holland, P.W. e Thayer, D.T. (1981). *Section pre-equating: the Graduate Record Examination*. Program Statistics Research Technical Report N° 81-13, Princeton, NJ: Educational Testing Service.

[66] Jiang, Y. (2005). *Estimating parameters for multidimensional item response theory models by MCMC methods*. Phd thesys. Department of Counselling, Educational Psychology and Special Education. East Lansing: Michigan State University.

[67] Johnson, R.A. e Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall.

[68] Johnson, R.A.; Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis* - 5th ed. New Jersey: Prentice Hall.

[69] Kim, S.H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement*, 25, 163-176.

[70] Kolen, M.J. e Brennan, R.L. (1995). *Test Equating: Methods and Practices*. New York: Springer.

[71] Kolen, M.J. e Brennan, R.L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices* - 2nd ed. New York: Springer.

[72] Kuder, G.F. e Richardson, M.W. (1937). The theory of the estimation of test reliability, *Psychometrika*, 2, 151-160.

[73] Lazarsfeld, P. (1966). Latent structure analysis. Em *Measurement and Prediction* (Stouffer, S.A.; Guttman, L.; Suchman, E.; Lazarfeld, P., Star, S. e Claussen, J., Eds.). New York: Wiley.

[74] Van der Linden, W.J. e Hambleton, R.K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer.

[75] Livingston, S.A. (2004). *Equating Test Scores (Without IRT)*. Princeton, NJ: Educational Testing Services.

[76] Lord, F.M. (1952). *A theory of test scores* - Nº 7. Psychometric Monograph.

[77] Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale NJ: Erlbaum.

[78] Lord, F.M. e Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. New York: Addison-Wesley.

[79] Loyd, B.H. e Hoover, H.D. (1980). Vertical equating using the Rasch Model. *Journal of Educational Measurement*, 17, 179-193.

[80] Loureiro, M.J.; Ferrão, M.E.; Navio, V.M.; Dias, V.M.; Tavares, A. e Teles, J. (2006). Avaliação do Autoconceito Infantil. *Actas da XI Conferência Internacional Avaliação Psicológica: Formas e Contextos*. Livro de Actas. Universidade do Minho.

[81] Marco, G.L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139-160.

[82] Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

[83] Matriz de Referência de Matemática do Projecto de Investigação: Eficácia Escolar no Ensino da Matemática (2005). Covilhã: Universidade da Beira Interior - Departamento de Matemática.

[84] McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometric Monograph*, n. 15, 1-167.

[85] McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

[86] Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H. e Teller, E. (1953). Equation of State Calculations by Fast Computing Machine. *Journal of Chemical Physics*, 21, 1087-1091.

[87] Muller, P. (1991). *Metropolis based posterior integration schemes*. Technical Report, Statistics Department, Purdue University.

[88] Muraki, E. (1992). A generalized partial credit model : Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.

[89] Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, 17, 351-363.

[90] Muraki, E. e Bock, R.D. (2002). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks* (Version 3) [Computer software]. Chicago: Scientific Software.

[91] Muraki, E. e Engelhard, G. (1985). Full Information Item Factor Analysis: applications of EAP scores. *Applied Psychological Measurement*, 9, 417-430.

[92] Patz, R.J. e Junker, B.W. (1999a). A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models. *Journal of Educational and Behavioral Statistics*, 24, n. 2, 146-178.

[93] Patz, R. J. e Junker, B. W. (1999b). Applications and Extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24, n. 4, 342-366.

[94] Petersen, N.S.; Kolen, M.J. e Hoover, H.D. (1989). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematicaland Statistical Psychology*, 40, 43-49.

[95] Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

[96] Rasch, G. (1962). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 4, 321-334.

[97] Reckase, M.D. (2009). *Mutidimensional Item Response Theory*. New York: Springer.

[98] Relatório 1: *Provas de Aferição de Matemática, Português do 4º e 6º anos de escolaridade* (2009). Relatório Técnico. Covilhã: Universidade da Beira Interior, Departamento de Matemática.

[99] Relatório 2: *Testes 3EMat e provas de aferição* (2009). Relatório Técnico. Covilhã: Universidade da Beira Interior, Departamento de Matemática.

[100] Rubin, D.B. (1980). Using empirical Bayes techniques in the Law School validity studies. *Journal of the American Statistical Association*, 75, 801-827.

[101] Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement*, n. 17.

[102] Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional space. *Psychometrika*, 39, 111-121.

[103] Samejima, F. e Livingston, P. (1979). *Method of moments as the least squares solution for fitting a polynomial* (Research Report 79-2). Knoxville: University of Tennessee, TN.

[104] Sánchez, A.V. e Escribano, E.A. (1999a). *Desarrollo y Evaluación del Autoconcepto en la Edad Infantil*. Bilbao: Ediciones Mensagero.

[105] Sánchez, A.V. e Escribano, E.A. (1999b). *Medição do auto-conceito*. São Paulo: Editora da Universidade do Sagrado Coração.

[106] Soares, T.M. (2005). Utilização da Teoria da Resposta ao Item na Produção de Indicadores Sócio-Econômicos. *Pesquisa Operacional*, 25, n. 1, 83-112, jan./abr. Rio de Janeiro.

[107] Soares, T.M.; Gonçalves, F.B. e Gamerman, D. (2009). An Integrated Bayesian Model For DIF Analysis. *Journal of Educational and Behavioral Statistics*, 3, 348-377.

[108] Stocking, M.L. e Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.

[109] Sympson, J.B. (1978). A model for testing with multidimensional items. Em *Proceedings of the 1977 Computerized Adaptative Testing Conference* (Weiss, D.J., Ed.). Minneapolis: University of Minnesota.

[110] Sympson, J.B. (1983). *A new IRT model for calibrating multiple choice items.* Paper presented at the annual meeting of the Psychometric Society , Los Angeles, CA.

[111] Thurstone, L.L. (1947). Multiple-factor analysis. Chicago: University of Chicago Press.

[112] Toit, M. (2003). *Irt from SSI: Bilog-mg, Multilog, Parscale, Testfact.* Lincolnwood, Illinois: Scientific Software International.

[113] De la Torre, J. e Patz, R. J. (2002). *A Multidimensional Item Response Theory Approach to Simultaneous Ability Estimation.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, United States of America: New Orleans.

[114] Vicente, P. (2007). Plano Amostral do Projecto 3EM - Eficácia Escolar no Ensino da Matemática. Em *Estatística: Ciência Interdisciplinar. Actas do XIV Congresso Anual da Sociedade Portuguesa de Estatística* (Ferrão, M.E.;

Nunes, C. e Brauman, C., Eds.). Covilhã: Edição da Sociedade Portuguesa de Estatística.

[115] Wilson, D.T.; Wood R. e Gibbons, R. (1998). *Testfact: Test Scoring, and Item Factor Analysis.* Lincolnwood, Illinois: Scientific Software International, Inc.

[116] Wollack, J.A.; Bolt, D.M.; Cohen, A.S. e Lee, Y-S. (2001). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, 26, 337-350.

[117] Wright, B.D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97-116.

[118] Zimowski, M.; Muraki, E.; Mislevy, R. e Bock, D. (2003). *Bilog-Mg 3 for Windows.* Chicago: Scientific Software International.

# Anexos

# Anexo 1

## Código do programa em Matlab para obter as estimativas dos parâmetros do MMRI logístico de 2 parâmetros

```
-----------------------------------------------------------------------------------------------
% PROGRAMA PRINCIPAL

dimest=2 % definição do número de dimensões do teste
% (Valores iniciais dos parâmetros dos itens)
a0b=a0;
a0b=ones(I,dimest);
b0 = zeros(I,1);

% (Valores iniciais dos factores latentes)
theta0 = normrnd(zeros(J,1),1);
theta0 = (theta0-mean(theta0))/std(theta0);
theta0b = repmat(theta0,1,dimest);

%(Variâncias consideradas a priori para os parâmetros dos itens e dos factores latentes)
var_loga_prior = 1
var_b_prior = 2
var_t_prior = 1;

%(Variância a priori)
cand_t_var = 0.01
cand_a_var = 0.01
cand_b_var = 0.01

%(Cria matrizes vazias para guardar os resultados das últimas iterações)
Mtheta_1=[];
Mtheta_2=[];
Mtheta=[];
Ma=[];
%Ma_1=[];
%Ma_2=[];
Mb=[];

%(Contador de iterações da cadeia - 15000 iterações)
for k = 1:15000
theta1b = FuncProf(theta0b,a0b,b0,I,J,y,var_t_prior,cand_t_var);
[b1] = FuncPar(theta1b,a0b,b0,I,J,y,var_loga_prior,var_b_prior,cand_a_var,cand_b_var);
[a1b] = FuncPara(theta1b,a0b,b0,I,J,y,var_loga_prior,var_b_prior,cand_a_var,cand_b_var);
```

```
theta0b = theta1b;
a0b = a1b;
b0 = b1;
k
     if k>14000
       Mtheta=[Mtheta, theta0b];
       Ma = [Ma a0b];
       Mb=[Mb b0];
     end
end
```

---

```
% Gera os dados para o MMRI de 2 parâmetros

nalunos=2000;
dim=2;
nitens=40;
sigma=diag(ones(dim,1));sigma(1,2)=0.3;sigma(2,1)=0.3;
tetav=mvnrnd(zeros(dim,1),sigma,nalunos);
x=1;
av=[0 1
   0 1
   0 1
   0 1
   0 1
   0 1
   0 1
   0 1
   0 1
   x x
   x x
   x x
   x x
   x x
   1 0
   1 0
   1 0
   1 0
   1 0];
av=[av;av];
bv=normrnd(0,1.4,nitens,1);
sumat=zeros(nalunos,nitens);
```

```
suma=zeros(1,nitens);
for i=1:dim,
    suma=suma+av(:,i)';
    sumat=sumat+(repmat(tetav(:,i),1,nitens).*repmat(av(:,i)',nalunos,1));
end
suma=repmat(suma.*bv',nalunos,1);
DEN=ones(nalunos,nitens)+exp(-sumat+suma);
PROB=ones(nalunos,nitens)./DEN;
dados=binornd(1,PROB);
a0=av;
y=dados;
J=nalunos;
I=nitens;
```

---

```
% Func2PL - MMRI logístico de 2 parâmetros

 function p = Func2PL(theta,a,b,I,J)
 z1 = (theta*a');
 z2 = sum(a,2).*b;
 z = -z1 + repmat(z2',J,1);
 p = (1./(1.+exp(z))).
```

---

```
% FuncProf  - Função que calcula as estimativas dos factores latentes
% (Esta funçao utiliza a funçao FuncVer_prof para o cálculo da verosimilhança dos factores latentes)

function theta1b = FuncProf(theta0b,a0b,b0,I,J,y,var_t_prior,cand_t_var)
[aa,dim]=size(theta0b);
theta1b = theta0b + normrnd(0,sqrt(cand_t_var),J,dim);
mu = zeros(1,dim);
sigma = eye(dim);
alphatheta_num1 = (FuncVer_prof(theta1b,a0b,b0,I,J,y)) + log(mvnpdf(theta1b,mu,sigma));
alphatheta_den1 = (FuncVer_prof(theta0b,a0b,b0,I,J,y)) + log(mvnpdf(theta0b,mu,sigma));
alpha_theta_aux1 = alphatheta_num1-alphatheta_den1;
alpha_theta1 = min(ones(J,1),exp(alpha_theta_aux1));
prob_theta1 = binornd(ones(J,1),alpha_theta1);
theta1_1 = repmat(prob_theta1,1,dim).*theta1b+repmat((1-prob_theta1),1,dim).*theta0b;
theta1b = theta1_1;
```

---

```
% FuncVer_prof - Função que calcula a verosimilhança para os factores latentes
% (Utiliza a função Func2PL para o cálculo da probabilidade do examinando acertar no item)
```

```
function V_prof = FuncVer_prof(theta,a,b,I,J,y)
L_FiV =  ((y==1).*(Func2PL(theta,a,b,I,J))) + ((y==0).*((ones(J,I)-Func2PL(theta,a,b,I,J))));
V_prof = sum(log(L_FiV'))';
```

---

```
% FuncPar - Funçao que calcula as estimativas do parâmetro de dificuldade dos itens
%(Utiliza a função FuncVer_par para o cálculo da verosimilhança)

function [b1] = FuncPar(theta0b,a0b,b0,I,J,y,var_loga_prior,var_b_prior,cand_a_var,cand_b_var)
la0 = log(a0b);
la1 = la0;
a1b = exp(la1);
b1 = b0 + normrnd(0,sqrt(cand_b_var),I,1);
alphapar_num1 = FuncVer_par(theta0b,a1b,b1,I,J,y) + log(normpdf(b1(:,1),0,sqrt(var_b_prior)));
alphapar_den1 = FuncVer_par(theta0b,a0b,b0,I,J,y) + log(normpdf(b0(:,1),0,sqrt(var_b_prior)));
alpha_par_aux1 = alphapar_num1-alphapar_den1;
alpha_par1 = min(1,exp(alpha_par_aux1));
prob_par1 = binornd(ones(I,1),alpha_par1);
b1 = prob_par1.*b1(:,1)+(1-prob_par1).*b0(:,1);
```

---

```
% FuncPara - Calcula as estimativas dos parâmetros de discriminação dos itens
% (Utiliza a função FuncVer_par para o cálculo da verosimilhança)

function [a1b] = FuncPara(theta0b,a0b,b0,I,J,y,var_loga_prior,var_b_prior,cand_a_var,cand_b_var)
[aa,dim]=size(theta0b);
la0 = (a0b);
if det(corrcoef(la0)) >0, sigma=diag(ones(dim,1)*0.01)*corrcoef(la0)*diag(ones(dim,1)*0.01);
else sigma=diag(ones(dim,1));
end
la1 = mvnrnd(la0,sigma);
sigmap=eye(dim);
mediap=zeros(1,dim);
a1b = (la1);
b1 = b0;
alphapar_num1 = FuncVer_par(theta0b,a1b,b0,I,J,y) + log(mvnpdf(la1,mediap,sigmap));
alphapar_den1 = FuncVer_par(theta0b,a0b,b0,I,J,y) + log(mvnpdf(la0,mediap,sigmap));
alpha_par_aux1 = alphapar_num1-alphapar_den1;
alpha_par1 = min(1,exp(alpha_par_aux1));
prob_par1 = binornd(ones(I,1),alpha_par1);
a1b = repmat(prob_par1,1,dim).*a1b+repmat((1-prob_par1),1,dim).*a0b;
```

---

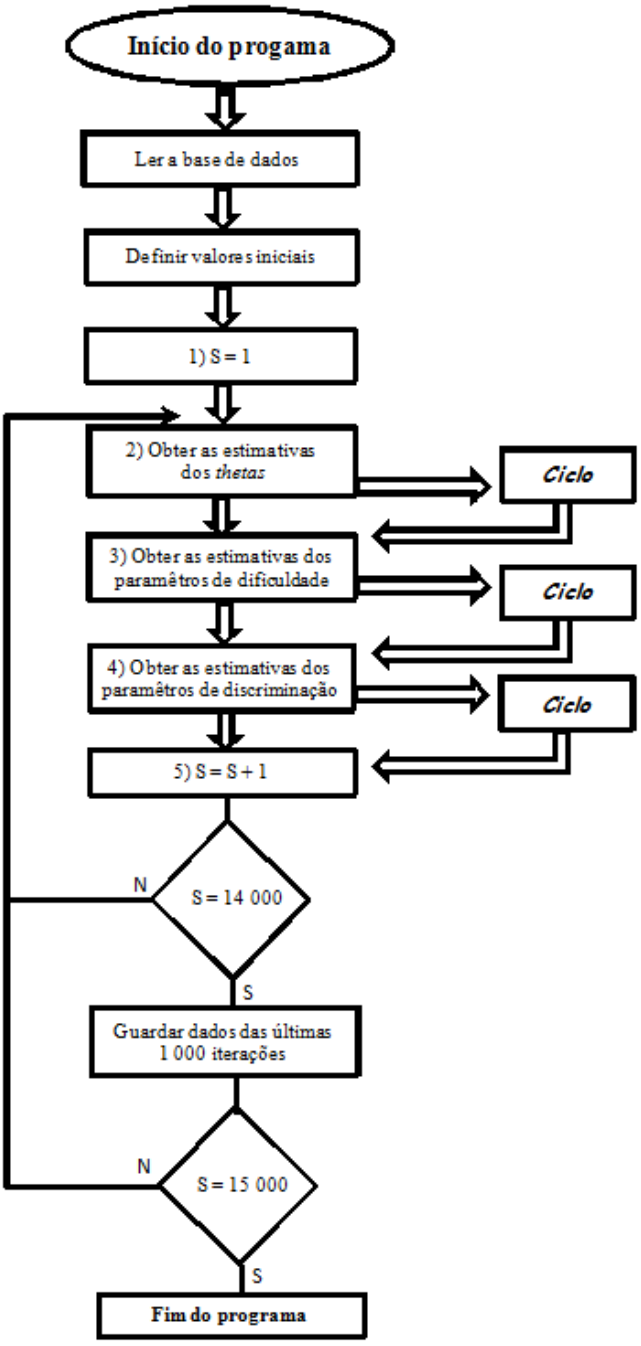
```
% FuncVer_par - Função que calcula a verosimilhança para os parâmetros dos itens
% (Utiliza a função Func2PL para o cálculo da probabilidade do examinando acertar no item)

function V_par = FuncVer_par(theta,a,b,I,J,y)
L_FiV = (y==1).*(Func2PL(theta,a,b,I,J))+(y==0).*((ones(J,I)-Func2PL(theta,a,b,I,J)));
V_par = sum(log(L_FiV))';
```

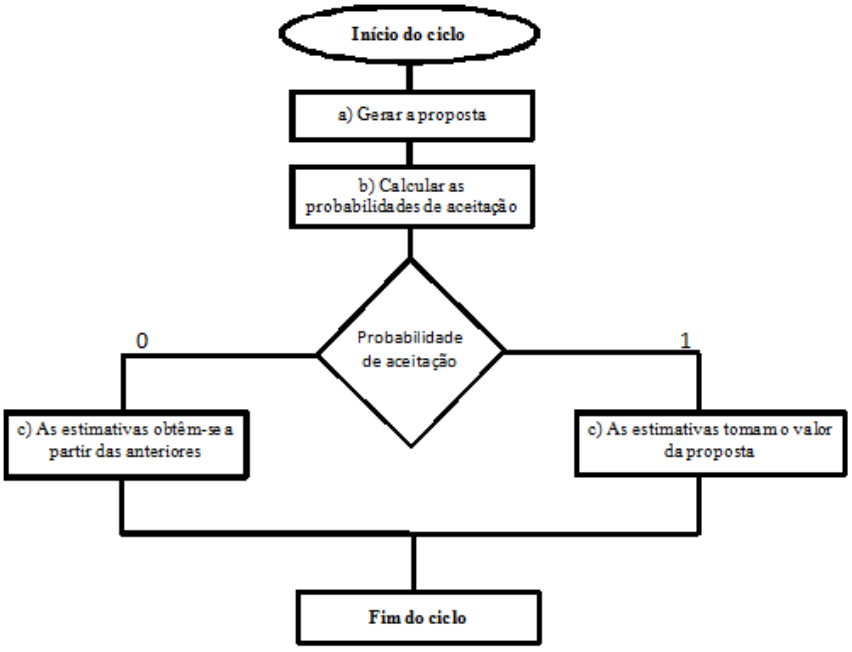
---

# Anexo 2

**Fluxograma do procedimento de estimação proposto para a estimação dos parâmetros do MMRI logístico de 2 parâmetros**



## Ciclo do fluxograma



# Anexo 3

## Descrição das bases de dados

### 3.1 - Dados de Estatística

Na Universidade do Minho está a decorrer um projecto que visa a utilização, de forma complementar, da TCT e dos MRI para garantir a qualidade na aferição das aprendizagens, na unidade curricular de Estatística do curso de Mestrado Integrado em Engenharia e Gestão Industrial. Os dados referem-se a três testes de escolha múltipla utilizados para aferir competências em Estatística, e foram aplicados aos alunos que frequentaram o mestrado nos anos lectivos 2006/2007 e 2007/2008. Os 1º e 2º testes aferem competências em Estatística Descritiva e Probabilidades, respectivamente. Estes instrumentos são compostos por 20 itens, no entanto, foram retirados, no 1º ano o item 9 e no 2º ano o item 11, uma vez que estes itens não foram respondidos por alunos de ambos os anos lectivos. Ao teste do 1º ano responderam 158 alunos e no do 2º ano, 161 alunos. O 3º teste afere competências em Distribuições, é constituído por 24 itens e foi respondido por 135 alunos. Foram estudadas as propriedades dos itens e a fiabilidade dos três testes. Para analisar a qualidade de itens/testes e comparar as classificações obtidas pelos alunos que frequentaram a unidade curricular nos dois anos lectivos, foi considerado cada conjunto de alunos como amostras independentes da mesma população. A análise dos instrumentos permitiu identificar os itens discriminativos e os níveis de dificuldade dos diversos itens, bem como, a consistência interna dos instrumentos, paralelamente com a função de informação dos itens e do teste. Apesar de as duas abordagens (TCT e MRI) coincidirem no que diz respeito aos achados principais, realça-se a importância da análise baseada em MRI, pela possibilidade da criação de bancos de itens e assim, os itens/testes utilizados na aferição das aprendizagens, permitirem a comparação temporal dos resultados (Costa, Oliveira e Ferrão [30]; Ferrão, Costa e Oliveira [45]).

A utilização de testes de múltipla escolha surge pela necessidade de adaptação a uma nova cultura de ensino-aprendizagem-avaliação no contexto da Declaração de Bolonha e, ao mesmo tempo, a adopção de novos métodos de avaliação do aluno que permitam garantir elevados padrões de qualidade como os proclamados pela Associação Europeia para a Garantia da Qualidade no Ensino Superior (ENQA [43]). Ainda no contexto de Bolonha, estes testes permitem fornecer aos alunos informação sobre o seu progresso ao longo da unidade curricular, constituindo um factor motivador das aprendizagens e, nesse sentido, influenciam o próprio processo de aprendizagem. Outra das vantagens da utilização deste tipo de testes de escolha múltipla é permitirem uma maior cobertura dos programas das unidades curriculares.

### 3.2 - Dados das Provas de Aferição de Matemática

O projecto Melhoria da Qualidade dos Instrumentos e Escalas de Aferição dos Resultados Escolares foi desenvolvido em parceria entre o Gabinete de Avaliação Educacional (GAVE) do Ministério da Educação, e a Universidade da Beira Interior (UBI), nos termos estabelecidos no Protocolo celebrado para o efeito. Os dados em análise foram recolhidos através da aplicação da prova de aferição de Matemática (PAM) do 4º ano de escolaridade do Ensino Básico, no final do ano lectivo 2006/2007. As provas de aferição do Ensino Básico são da responsabilidade do GAVE [50] e visam avaliar o modo como os objectivos e as competências essenciais de cada ciclo estão a ser alcançados pelo sistema de ensino. Estas provas são aplicadas anualmente a todos os alunos matriculados no quarto e sexto anos de escolaridade, em concordância com o disposto no Despacho n.º 2351/2007, de 14 de Fevereiro, Série II. A aferição, de acordo com o consignado no art.º17.º do Decreto-Lei n.º 6/2001, de 18 de Janeiro, visa a recolha de dados relevantes sobre os níveis de desempenho dos alunos, no que respeita às aprendizagens adquiridas e competências desenvolvidas. As provas, enquanto instrumentos de aferição, fornecem indicadores das aprendizagens dos alunos, tendo por referência as competências específicas da

disciplina de Matemática apresentadas no Currículo Nacional do Ensino Básico - Competências Essenciais e o programa em vigor. Em particular, a PAM tem como propósito avaliar: a compreensão de conceitos e procedimentos, a capacidade de raciocínio e de comunicação e a competência para usar a Matemática na análise e resolução de problemas.

As provas de aferição de 1º ciclo têm a duração de 90 minutos, repartidos por dois períodos de 45 minutos, separados por um intervalo de 25 minutos. As provas de aferição de 2º ciclo têm a duração de 100 minutos, repartidos por dois períodos de 50 minutos, separados por um intervalo de 20 minutos. As PAM de 4º e 6º anos (daqui em diante designadas por PAM4 e PAM6) são compostas por duas partes idênticas que incluem as seguintes áreas temáticas: números e cálculo; geometria e medida; estatística e probabilidades; e álgebra e funções. Ambas as provas são constituídas por 27 itens, contendo itens de resposta curta, de escolha múltipla, de completamento e de resposta aberta.

Os dados da PAM4 e PAM6 referem-se ao ano lectivo 2006/2007.

### 3.3 - Dados do Projecto de 3EM

O projecto de investigação Eficácia Escolar no Ensino da Matemática - 3EM é da responsabilidade da UBI, concretamente do Departamento de Matemática e do Departamento de Psicologia e Educação. Decorreu entre 2004 e 2008, com o co-financiamento do Ministério da Ciência, Tecnologia e Ensino Superior e da Fundação Calouste Gulbenkian. Este projecto foi coordenado pela Professora Doutora Maria Eugénia Ferrão e seguiu metodologicamente a linha de investigação em Eficácia Escolar. Teve como principais objectivos: a estimação do efeito-escola e a identificação dos factores intra-escolares que contribuem para a melhoria da qualidade da Educação em termos dos resultados escolares em Matemática e do processo ensino-aprendizagem. Tratou-se de um estudo longitudinal, no qual a recolha de dados se realizou no início e no final do ano lectivo, e envolveu dois coortes de alunos, dos 1º,

2º e 3º ciclos do Ensino Básico da região da Cova da Beira (concelhos de Covilhã, Fundão e Belmonte). Em particular, os níveis de ensino envolvidos no estudo no ano lectivo 2005/6, e que compõem o 1º coorte de alunos, foram os correspondentes a 1º, 3º, 5º, 7º e 8º anos de escolaridade. No ano lectivo 2006/7 procedeu-se ao acompanhamento destes alunos e entrou para o estudo um novo coorte da 1º, 3º, 5º e 7º anos. Todos estes alunos foram acompanhados no ano lectivo 2007/8. Os testes aplicados para aferir aprendizagens a Matemática designam-se 3EMat (Ferrão *et al.* [44]) e foram construídos a partir da Matriz de Referência de Matemática do projecto de investigação 3EM [83]. Estes testes foram aplicados a uma amostra aleatória (Vicente [114]) de alunos do Ensino Básico da região da Cova da Beira. Na composição de cada teste foi assegurada a representatividade de todos os conteúdos programáticos previstos no currículo nacional, tendo em conta o peso de cada um e a complexidade das tarefas propostas. Os testes 3EMat são constituídos por itens de múltipla escolha, com quatro opções de resposta, retirados de um banco de itens construído para medir as competências desenvolvidas em Matemática no Ensino Básico. Cada teste continha itens do ano lectivo anterior, com vista, à futura criação de uma escala vertical de desempenho a Matemática. Mais detalhes sobre a metodologia adoptada para o desenvolvimento dos instrumentos 3EMat, em particular, no que se refere à constituição de banco de itens e à criação do teste a partir do banco de itens, podem ser encontrados em Ferrão *et al.* [44]. A aplicação dos testes do 1º ano foi feita individualmente e a aplicação dos testes dos restantes anos de escolaridade foi efectuada colectivamente. Cada teste aplicado nos anos de escolaridade correspondentes aos 1º e 2º ciclos era composto por 30 itens e os testes aplicados a alunos do 3º ciclo continham 33 itens. Os testes para os alunos do 3º ciclo estavam divididos em duas partes, a primeira parte era realizada sem recorrer à calculadora e a segunda parte era efectuada com a ajuda de calculadora.

No âmbito do projecto 3EM foram aplicados outros instrumentos. O intrumento que avalia a Percepção do Autoconceito Infantil - PAI foi desenvolvido por Sánchez

e Escribano ([104] e [105]), validado para a população portuguesa (Loureiro *et al.* [80]). Foi aplicado no início do ano lectivo 2005/2006 e apresenta uma forma para examinandos do sexo masculino e outra forma para examinandos do sexo feminino. Os itens são de resposta dicotómica, constituídos por figuras, em que qualquer dos examinandos, menino ou menina, realiza uma actividade que se pode considerar representativa de um autoconceito positivo e representativa de um autoconceito negativo. A aplicação deste instrumento foi realizada colectivamente, em pequenos grupos de alunos e era composta por uma formulação oral dos itens, feita pelo aplicador na sala de aula, e por uma representação gráfica destinada aos examinandos avaliados (Loureiro *et al.* [80]).

# Anexo 4

## Prova de Aferição de Matemática do 4º ano

Prova de Aferição de Matemática – 1.º Ciclo do Ensino Básico 2007

A preencher pelo Aluno

Nome:

A preencher pela U.E.

N.º convencional do aluno: N.º convencional da escola:

N.º convencional do aluno: N.º convencional da escola:

2007

Prova de Aferição de Matemática

1.º Ciclo do Ensino Básico

## Instruções Gerais sobre a Prova

- A prova deve ser realizada a lápis.
- Podes usar borracha, apara-lápis e régua graduada.
- Se precisares de alterar alguma resposta, apaga-a e escreve a nova resposta.
- Em algumas questões, terás de colocar **X** no quadrado correspondente à resposta correcta. Se te enganares e puseres **X** no quadrado errado, apaga-o e volta a colocar **X** no lugar que consideres certo.
- Não apagues as contas e os desenhos que utilizares nas tuas respostas.
- Responde a todas as perguntas com a máxima atenção.
- Se acabares antes do tempo previsto, deves aproveitar para rever a tua prova.

**A prova tem duas partes.**

**No fim da Primeira Parte há um intervalo.**

**Tens 45 minutos para responder a cada parte.**

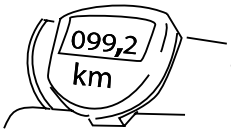
# Parte A

**1.** O Nuno e a Clara combinaram ir à gelataria de bicicleta.

**1.1.** Da casa do Nuno à casa da Clara são **0,4 km**.
Qual é a distância, em **metros**, da casa do Nuno à casa da Clara?

Resposta: _____ m

**1.2.** Vê, na imagem, quantos quilómetros marcava o conta-quilómetros da bicicleta da Clara quando ela saiu de casa.




A gelataria fica a **1 km de distância** da casa da Clara.

Assinala com **X** o valor que marcava o conta-quilómetros da bicicleta da Clara quando ela chegou à gelataria.

☐ 99,1

☐ 99,3

☐ 100,2

☐ 101,2

**2.** Na gelataria, há quatro ingredientes diferentes para colocar por cima do gelado:

- raspa de chocolate;
- amêndoa;
- chantili;
- gomas.

O Nuno escolheu um gelado e decorou-o colocando por cima raspa de chocolate e amêndoa.
A Clara também queria decorar o seu gelado com **dois** dos ingredientes.

De quantas maneiras diferentes poderia a Clara decorar o seu gelado?

Explica como chegaste à tua resposta. Podes fazê-lo utilizando palavras, desenhos ou contas.

Resposta: _____

**3.** Assinala com **X** os **sólidos** que o gelado do Nuno te faz lembrar.

☐ um círculo e uma pirâmide.

☐ um cone e uma esfera.

☐ um paralelepípedo e um cone.

☐ um triângulo e uma esfera.

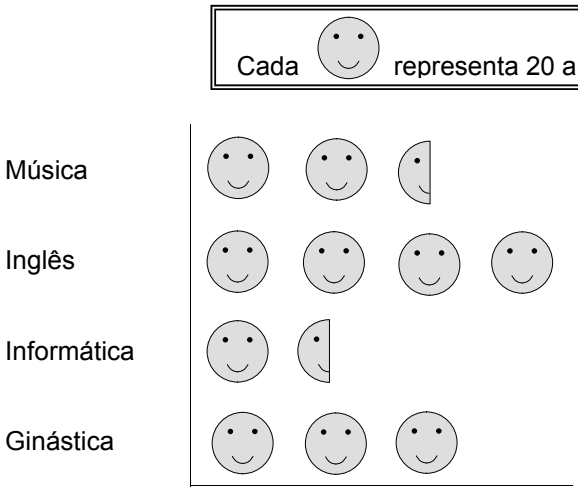
**4.** A Clara pediu na gelataria a receita do gelado.
Deram-lhe a seguinte receita para 4 pessoas.

| **Receita para 4 pessoas** | |
|---|---|
| morangos | 250 g |
| açúcar | 140 g |
| natas | 6 dl |
| limão | 1 |

Completa a receita que a Clara deve usar para fazer o gelado para 8 pessoas.

| **Receita para 8 pessoas** | |
|---|---|
| morangos | 500 g |
| açúcar | _____ g |
| natas | _____ dl |
| limão | 2 |

**5.** Na escola do Nuno, depois das aulas, os alunos frequentam uma das actividades: Música, Inglês, Informática e Ginástica. A figura mostra como todos os alunos são distribuídos pelas quatro actividades, à 4ª feira.




**5.1.** Quantos alunos têm Informática à 4ª feira?

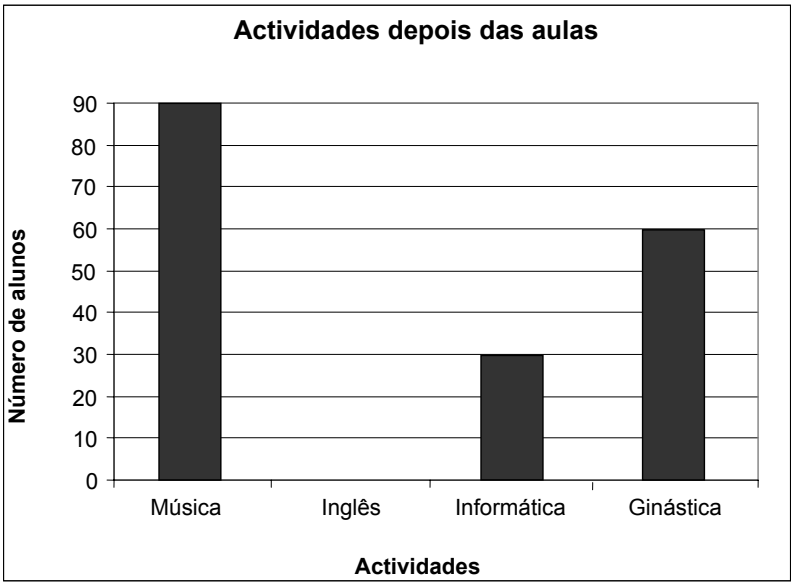
Resposta: _____

**5.2.** Escreve mais uma pergunta que possa ser respondida com informação da mesma figura.

_____

_____

**5.3.** O gráfico seguinte mostra como os mesmos alunos são distribuídos por três actividades à 5ª feira.




Calcula o número de alunos que têm Inglês à 5ª feira e desenha, no gráfico, a barra correspondente a esse número.

Explica o que fizeste para saberes quantos alunos têm Inglês à 5ª feira.

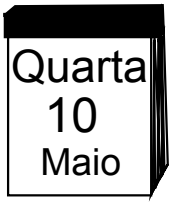
**6.** Observa o horário da turma do David, que está no 5º ano. Neste horário, podes ver as horas a que ele tem as aulas das diferentes disciplinas.

<table>
<tr><th>Horas</th><th>2ª feira</th><th>3ª feira</th><th>4ª feira</th><th>5ª feira</th><th>6ª feira</th></tr>
<tr><td>10:50 – 11:35</td><td rowspan="2">Educação Física</td><td></td><td></td><td></td><td></td></tr>
<tr><td>11:35 – 12:20</td><td></td><td></td><td></td><td>Educação Física</td></tr>
<tr><td>12:20 – 13:25</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>13:25 – 14:10</td><td>Formação Cívica</td><td>Estudo Acompanhado</td><td rowspan="2">Educação Visual e Tecnológica</td><td rowspan="2">Educação Musical</td><td rowspan="2">Matemática</td></tr>
<tr><td>14:10 – 14:55</td><td>História e Geografia de Portugal</td><td>Estudo Acompanhado</td></tr>
<tr><td>15:05 – 15:50</td><td rowspan="2">Inglês</td><td rowspan="2">Educação Visual e Tecnológica</td><td rowspan="2">História e Geografia de Portugal</td><td rowspan="2">Inglês</td><td rowspan="2">Língua Portuguesa</td></tr>
<tr><td>15:50 – 16:35</td></tr>
<tr><td>16:55 – 17:40</td><td>Ciências da Natureza</td><td rowspan="2">Área de Projecto</td><td rowspan="2">Língua Portuguesa</td><td rowspan="2">Matemática</td><td rowspan="2">Ciências da Natureza</td></tr>
<tr><td>17:40 – 18:25</td><td>Língua Portuguesa</td></tr>
</table>

**6.1.** Na **segunda-feira**, a que horas e minutos teve início a aula de Educação Física?

Resposta: _____horas e_____ minutos

**6.2.** A Clara olhou para o relógio e para o calendário que estão pendurados na cozinha.
Pensou na disciplina que o seu irmão David estava a ter àquela hora.

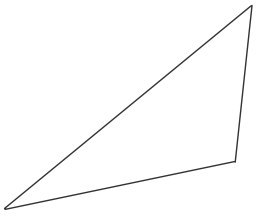

Escreve o nome da disciplina que ele teve nesse dia e a essa hora.

Resposta: _____

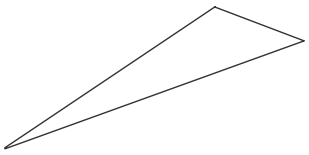
---

**7.** Um triângulo **isósceles** tem dois lados com igual comprimento.

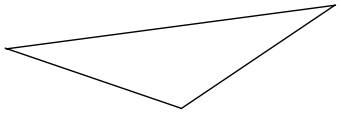
Assinala com **X** o triângulo que é isósceles.



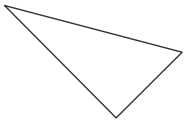
☐ Triângulo A



☐ Triângulo B

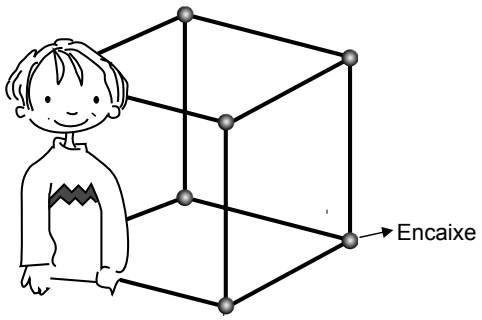


☐ Triângulo C



☐ Triângulo D

**8.** O Nuno utilizou um tubo de plástico para construir a estrutura de um cubo.




Cortou o tubo em bocados iguais, com 1 metro de comprimento cada um.

**8.1.** Para unir os tubos uns aos outros, o Nuno usou encaixes. Quantos encaixes usou?

Resposta: _____

**8.2.** Quantos metros de tubo utilizou o Nuno na sua construção?

Resposta: _____m


AQUI!

Não avances na prova até
o professor dizer.

Se acabaste antes do tempo previsto,
deves aproveitar para rever a tua prova.

# Parte B

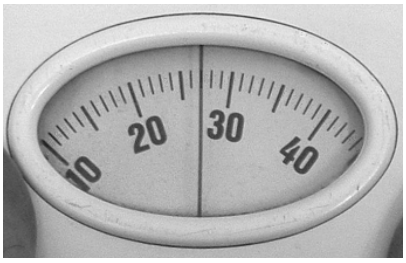
**9.** Calcula

31 + 719

Explica como chegaste à tua resposta. Podes fazê-lo utilizando palavras ou contas.

Resposta: _____

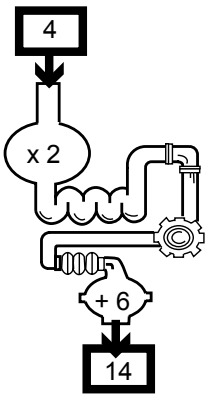
**10.** A balança mostra o peso do Nuno, em quilogramas.




Quanto pesa o Nuno, em quilogramas?

Resposta: _____kg

**11.** O Nuno leu a história do professor *Matema*, que construiu uma máquina dos números.
Quando o professor *Matema* colocou na abertura da máquina o número 4, saiu o 14.




**11.1.** Se o professor colocar na máquina o 12, que número sairá?

Resposta: _____

**11.2.** Que número teria de colocar na máquina o professor *Matema*, para que lhe saísse o 46?

Explica como chegaste à tua resposta. Podes fazê-lo utilizando palavras, desenhos ou contas.

Resposta: _____

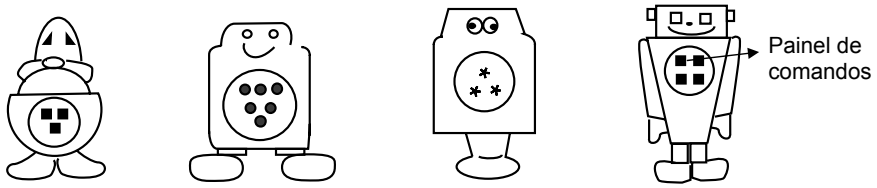
**12.** O livro da história do professor *Matema* conta que, um dia, ele construiu quatro robôs, o *Nume*, o *Reve*, o *Tal* e o *Zás*, de tal forma que:

— o *Zás* tem olhos quadrados;

— o *Reve* e o *Tal* não têm boca;

— o *Reve* não tem quadrados no painel de comandos.

Escreve, na linha por baixo de cada um dos robôs, o seu nome.

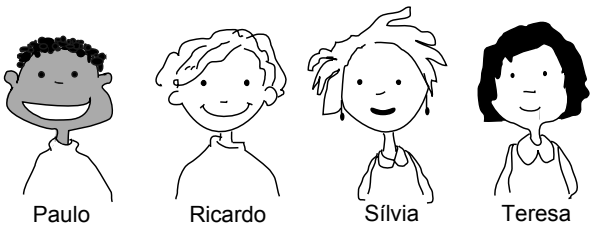



_____ _____ _____ _____

**13.** Completa a igualdade seguinte.

3872 − ☐ = 3072

**14.** Na figura, estão representados quatro amigos da Clara, que vivem em **países da Europa** e a quem ela enviou postais a desejar uma boa Páscoa.



Paulo **Espanha** | Ricardo **Espanha** | Sílvia **França** | Teresa **França**

Os quatro postais que a Clara comprou custaram, ao todo, **2 euros**. Mas ela ainda tinha os selos para comprar.

Observa a tabela de preços dos selos.

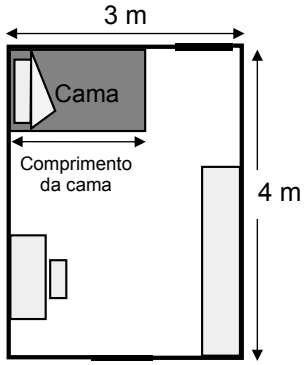
| | **Para Espanha** | **Para outros países da Europa** | **Para o resto do Mundo** |
|---|---|---|---|
| **Preços dos selos** | € 0,48 | € 0,57 | € 0,74 |

O que foi mais caro, os quatro selos ou os quatro postais?

Explica como chegaste à tua resposta. Podes fazê-lo utilizando palavras ou contas.

Resposta: _____

**15.** A planta seguinte é a do quarto da Clara.




**15.1.** Qual é a área do quarto da Clara, em metros quadrados?

Resposta: _____ $m^2$

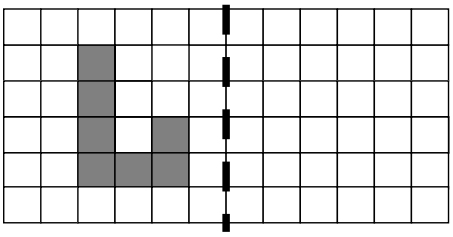
**15.2.** A cama da Clara tem 1 metro de largura.
Assinala com **X** o **comprimento**, aproximado, da cama da Clara.

- ☐ 0,75 metros
- ☐ 1,05 metros
- ☐ 1,75 metros
- ☐ 3,05 metros

**16.** Escreve **um número** que seja maior do que 7,8 e menor do que 8.

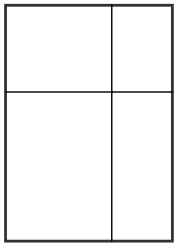
Número: _____

**17.** Desenha a figura simétrica da figura representada no quadriculado, relativamente ao eixo de simetria, indicado a tracejado.

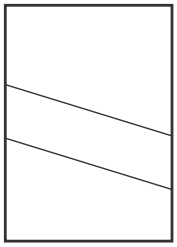


**18.** A professora da turma da Clara disse aos alunos que dobrassem uma folha de papel duas vezes, de modo a fazerem dois vincos na folha.
A Clara disse aos seus colegas de grupo:
— Olhem, os vincos da minha folha são **paralelos**.

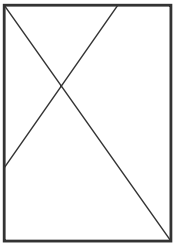
Assinala com **X** a figura que representa a folha da Clara.



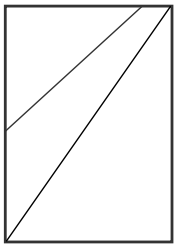
☐ Folha A



☐ Folha B

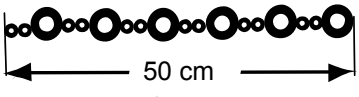


☐ Folha C



☐ Folha D

**19.** A Clara está a fazer um cinto com argolas grandes e argolas pequenas.
Já fez 50 cm do cinto, mas quer que o cinto tenha **2 metros** de comprimento, mantendo a sequência das argolas grandes e pequenas.




Com quantas argolas grandes e com quantas argolas pequenas ficará o cinto?

Explica como chegaste à tua resposta. Podes fazê-lo utilizando palavras, desenhos ou contas.

Argolas grandes: _____

Argolas pequenas: _____

**20.** Em Lisboa, vivem cerca de **560 mil** pessoas e, em Alcochete, vivem cerca de **10 000** pessoas.
Assinala com **X** o número, aproximado, de pessoas que vivem nestas duas localidades.

☐ 660 000

☐ 570 000

☐ 66 000

☐ 10 560

# Anexo 5

## Prova de Aferição de Matemática do 6º ano


Prova de Aferição de Matemática – 2.º Ciclo do Ensino Básico 2007

A preencher pelo Aluno

Nome:

A preencher pela U.E.

N.º convencional do aluno: N.º convencional da escola:

N.º convencional do aluno: N.º convencional da escola:

2007

Prova de Aferição de Matemática

2.º Ciclo do Ensino Básico

## Instruções Gerais sobre a Prova

- A prova deve ser realizada com caneta ou esferográfica de tinta azul ou preta, com excepção das questões em que te é indicado que resolvas a lápis.
- Podes usar borracha, apara-lápis, régua graduada e calculadora, mas não podes usar transferidor.
- Se precisares de alterar alguma resposta, risca-a e escreve a nova resposta.
- Em algumas questões, terás de colocar **X** no quadrado correspondente à resposta correcta. Se te enganares e puseres **X** no quadrado errado, risca esse quadrado e volta a colocar **X** no lugar que consideras certo.
- Não risques os cálculos e/ou os esquemas que utilizares nas tuas respostas.
- Responde a todas as perguntas com o máximo de atenção.
- Se acabares antes do tempo previsto, deves aproveitar para rever a tua prova.

**A prova consta de duas partes.**

**No fim da Primeira Parte, há um intervalo.**

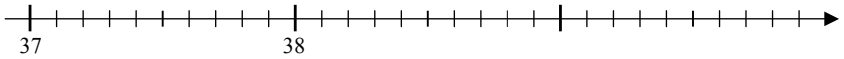
**Tens 50 minutos para responder a cada parte.**

# Parte A

**1.** A Bela está doente. Durante o dia, mediu quatro vezes a sua temperatura, com um termómetro.
Na tabela, estão representadas as temperaturas e as horas a que foram medidas.

| Horas | 8 | 12 | 16 | 20 |
|---|---|---|---|---|
| Temperatura (em ºC) | 38,5 | 38,9 | 39,2 | 38,7 |

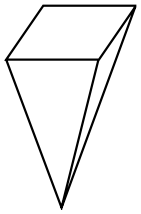
**1.1.** Assinala na recta numérica, com **X**, os pontos que correspondem às temperaturas registadas na tabela.




**1.2.** Qual é a diferença entre a temperatura registada às 16 horas e a registada às 20 horas?

Resposta: _____ºC

**2.** Quantos vértices, arestas e faces tem uma pirâmide quadrangular?



**2.1.** Número de vértices: _____

**2.2.** Número de arestas: _____

**2.3.** Número de faces: _____

**3.** O Gil comprou amêndoas da Páscoa, umas eram azuis e outras brancas. As amêndoas compradas pelo Gil estão representadas na figura.

**Dois terços** das amêndoas que comprou eram azuis.
Quantas amêndoas azuis comprou o Gil?

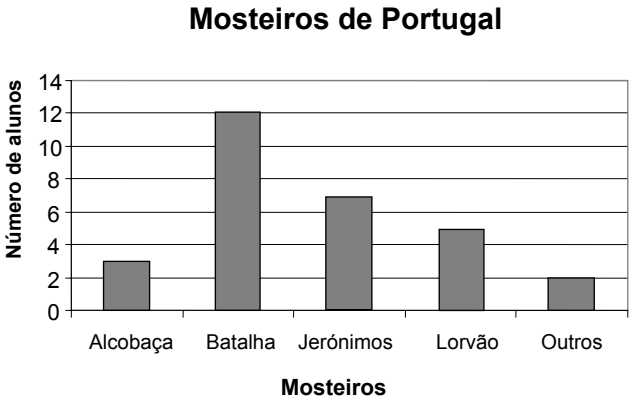
Resposta: _____

**4.** Na turma da Bela, todos os alunos responderam à questão:

«Que mosteiro de Portugal gostarias de visitar?»

Cada aluno deu uma única resposta.

Com as respostas obtidas, construíram o gráfico seguinte.




**4.1.** Quantos alunos tem a turma da Bela?

Resposta: _____

**4.2.** Escreve mais uma pergunta que possa ser respondida com informação do mesmo gráfico.

_____

_____

**5.** A turma do Gil foi visitar um mosteiro. À entrada, estavam dois cartazes: um com o preço dos bilhetes e outro com o número de visitantes do mosteiro.

| **Tipo de bilhetes** | **Preço** |
|---|---|
| Menos de 14 anos | (Gratuito) |
| Dos 14 aos 65 anos | 4 euros |
| Mais de 65 anos | 2 euros |

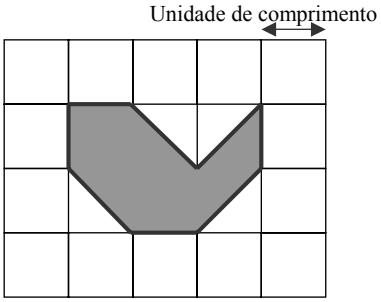
| | **Número de visitantes** | | |
|---|---|---|---|
| **Mês** / **Idade dos visitantes** | Janeiro | Fevereiro | Março |
| Menos de 14 anos | 500 | 850 | 750 |
| Dos 14 aos 65 anos | 300 | 150 | 250 |
| Mais de 65 anos | 50 | 50 | 100 |

Em qual dos três meses é que o mosteiro recebeu mais dinheiro pelos bilhetes vendidos?

Explica como chegaste à tua resposta. Podes fazê-lo utilizando palavras, esquemas ou cálculos.

Resposta: _____

**6.** Observa a figura desenhada no quadriculado.

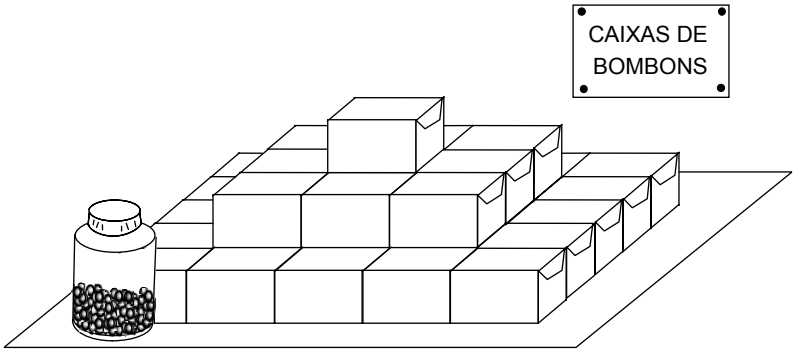



Assinala com **X** a frase que traduz uma afirmação verdadeira.

- ☐ O **perímetro** da figura é menor do que 4 unidades de comprimento.
- ☐ O **perímetro** da figura é igual a 4 unidades de comprimento.
- ☐ O **perímetro** da figura é igual a 8 unidades de comprimento.
- ☐ O **perímetro** da figura é maior do que 8 unidades de comprimento.

**7.** Uma das empregadas da loja de doces colocou várias caixas iguais umas sobre as outras, formando um monte como o que vês na figura.

O preço de uma caixa é de 1,78 euros.




Quanto paga um cliente por todas as caixas do monte?

Explica como chegaste à tua resposta. Podes fazê-lo utilizando palavras, esquemas ou cálculos.

Resposta: _____

**8.** Um rectângulo é um quadrilátero com quatro ângulos rectos.

***Um quadrado é um rectângulo, mas há rectângulos que não são quadrados***.

Tendo em conta as propriedades dos quadrados e as dos rectângulos, explica por que razão a frase anterior é verdadeira.

_____

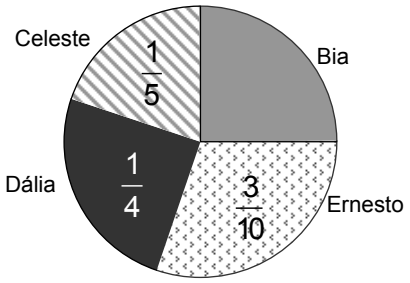
_____

_____

_____

**9.** Os quatro empregados da loja de doces, a Bia, a Celeste, a Dália e o Ernesto, arrumaram todos os chocolates nas prateleiras.

O gráfico refere-se à porção de chocolates que cada empregado arrumou.



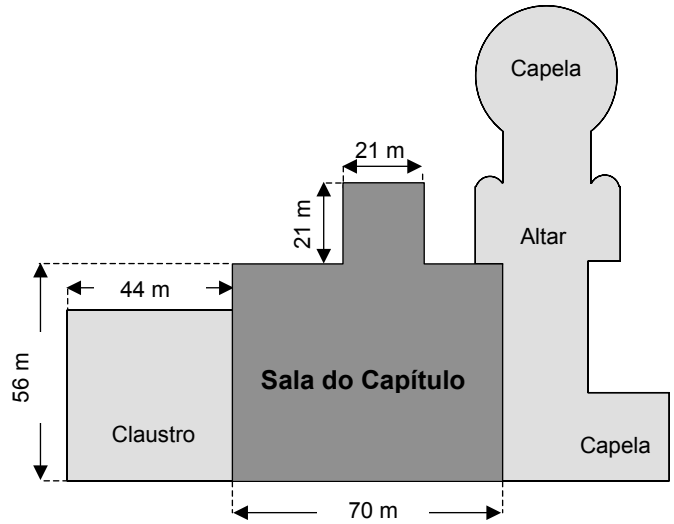

**9.1.** Que **percentagem** de chocolates arrumou o Ernesto?

Resposta: _____

**9.2.** Que **fracção** de chocolates arrumou a Bia?

Resposta: _____

**10.** Na figura, está representada a planta de um mosteiro.




De acordo com os comprimentos indicados na figura, calcula, em metros quadrados, a **área** da **Sala do Capítulo**.

Explica como chegaste à tua resposta. Podes fazê-lo utilizando palavras, esquemas ou cálculos.

Resposta: _____ $m^2$


AQUI!

Não avances na prova até
o professor dizer.

Se acabaste antes do tempo previsto,
deves aproveitar para rever a tua prova.

# Parte B

**11.** Na loja de doces, deram à Ana uma receita de gelado para **seis** pessoas.

Na tabela seguinte, estão as quantidades de cada um dos ingredientes da receita.

| **Receita para 6 pessoas** | |
|---|---|
| **Ingredientes** | **Quantidades** |
| ovos | 6 |
| açúcar | 1 chávena |
| leite com chocolate | 6 chávenas |
| baunilha | 3 colheres de café |
| chocolate preto | $\frac{1}{2}$ tablete |

Completa a tabela seguinte com as quantidades de ingredientes que a Ana deve usar ao fazer o gelado só para **três** pessoas.

| **Receita para 3 pessoas** | |
|---|---|
| **Ingredientes** | **Quantidades** |
| ovos | 3 |
| açúcar | _____ chávena |
| leite com chocolate | 3 chávenas |
| baunilha | _____colheres de café |
| chocolate preto | _____ tablete |

**12.** A Ana, o Gil, o Ivo e a Bela decidiram fazer uma maqueta de um mosteiro. Cada um deu 3 euros para comprar os materiais necessários.

A figura mostra as moedas que sobraram, depois de pagos todos os materiais.

Os quatro amigos distribuíram as moedas entre si, de modo a ficarem com iguais quantias de dinheiro.

Completa a tabela com o **número** de moedas de cada tipo que cada amigo recebeu. Repara que, na tabela, já foram distribuídas uma moeda de 1 euro e duas de 50 cêntimos.

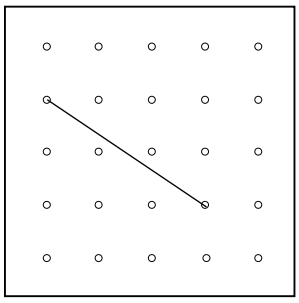
Utiliza o teu lápis para completares a tabela.

| | **Tipo de moedas** | | | | | | |
|---|---|---|---|---|---|---|---|
| |  |  |  |  |  |  |  |
| Ana | **1** | | | | | | |
| Gil | | **2** | | | | | |
| Ivo | | | | | | | |
| Bela | | | | | | | |

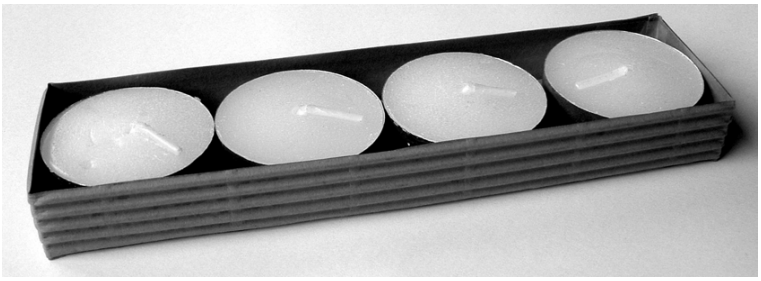
**13.** Escreve no ☐ um número, para completares a igualdade.

☐ : 4 = 3,1

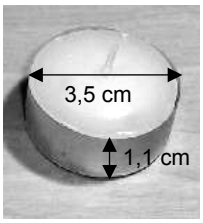
**14.** Na figura, está representada uma das diagonais de um rectângulo.
Desenha o rectângulo, utilizando o lápis e a régua.



**15.** A Ana comprou uma caixa de 4 velas, como a da figura.



Cada vela tem a forma de um cilindro com 1,1 cm de altura e 3,5 cm de diâmetro.

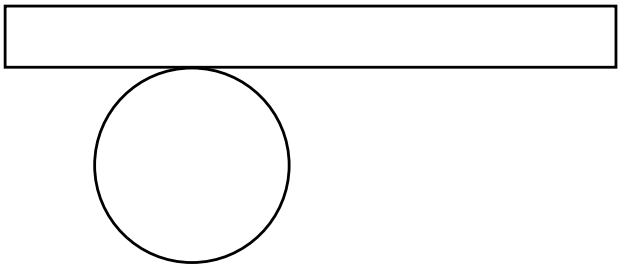



**15.1.** Determina, em $cm^3$, o volume aproximado da **caixa de quatro velas** que a Ana comprou.

Explica como chegaste à tua resposta. Podes fazê-lo utilizando palavras, esquemas ou cálculos.

Resposta: _____$cm^3$

**15.2.** A figura é uma planificação, em tamanho real, da tacinha de alumínio em que está contida uma das velas.



Qual é, aproximadamente, em centímetros, o perímetro do círculo da planificação?

Resposta: _____ cm

---

**16.** Calcula o valor da expressão numérica e apresenta os cálculos que efectuares.

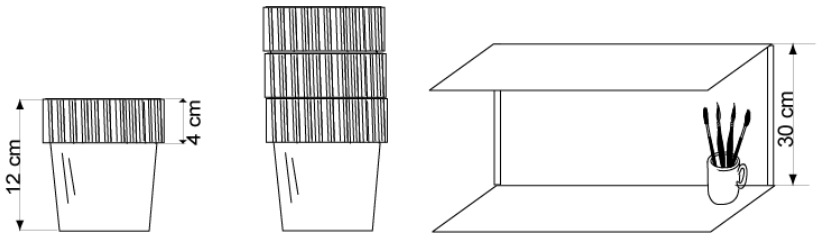
$$\frac{3}{5} + \frac{1}{2} : 0{,}4$$

Resposta: _____

**17.** Na sala de aula, há copos para os alunos lavarem os pincéis.

Cada copo tem 12 cm de altura e um rebordo com 4 cm.

A professora costuma guardar os copos numa prateleira. Para ocuparem menos espaço, encaixa-os uns nos outros, formando pilhas que não podem ultrapassar 30 cm de altura.



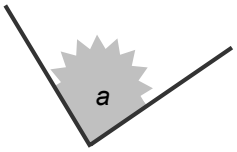

Copos Prateleira

No **máximo**, quantos copos pode ter cada pilha?

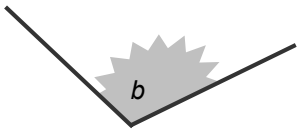
Explica como chegaste à tua resposta. Podes fazê-lo utilizando palavras, esquemas ou cálculos.

Resposta: _____

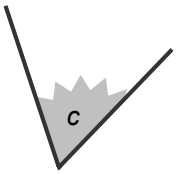
**18.** Assinala com **X** o ângulo que tem de amplitude mais de 120º e menos de 180º.



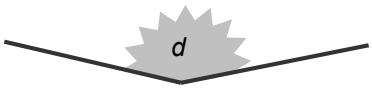

☐ Ângulo *a*




☐ Ângulo *b*




☐ Ângulo *c*

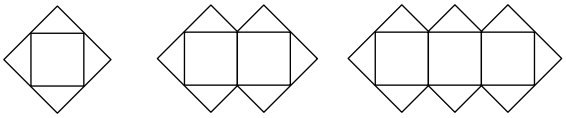



☐ Ângulo *d*

**19.** Escreve **um número** que seja, simultaneamente, múltiplo de 2, 3 e 5.

Número: _____

**20.** Observa a seguinte sequência de figuras.



1ª figura 2ª figura 3ª figura ...

**20.1.** Quantos triângulos terá a 5ª figura da sequência?

Resposta: _____

**20.2.** Desenha, utilizando o lápis e a régua, os eixos de simetria da figura representada a seguir.